UNIVERSITY OF ALBERTA

Application of Empirical Likelihood Estimation in Survival Data Analysis and Survey Sampling

BY

EMMANUEL BENHIN

A THESIS
SUBMITTED TO THE FACULTY OF GRADUATE STUDIES AND RESEARCH
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE
OF MASTER OF SCIENCE

DEPARTMENT OF STATISTICS AND APPLIED PROBABILITY

EDMONTON, ALBERTA
FALL 1994

National Library of Canada
Acquisitions and Bibliographic Services Branch
395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale du Canada
Direction des acquisitions et des services bibliographiques
395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Your file Votre référence

Our file Notre référence

The author has granted an irrevocable non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of his/her thesis by any means and in any form or format, making this thesis available to interested persons.

L'auteur a accordé une licence irrévocable et non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de sa thèse de quelque manière et sous quelque forme que ce soit pour mettre des exemplaires de cette thèse à la disposition des personnes intéressées.

The author retains ownership of the copyright in his/her thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without his/her permission.

L'auteur conserve la propriété du droit d'auteur qui protège sa thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

ISBN 0-315-95005-6

Canada

Name EMMANUEL BENHIN

Dissertation Abstracts International is arranged by broad, general subject categories. Please select the one subject which most nearly describes the content of your dissertation. Enter the corresponding four-digit code in the spaces provided.

STATISTICS

SUBJECT TERM

0463

SUBJECT CODE

U·M·I

Subject Categories

THE HUMANITIES AND SOCIAL SCIENCES

COMMUNICATIONS AND THE ARTS

- Architecture 0729
- Art History 0377
- Cinema 0900
- Dance 0378
- Fine Arts 0357
- Information Science 0723
- Journalism 0391
- Library Science 0399
- Mass Communications 0708
- Music 0413
- Speech Communication 0459
- Theater 0465

EDUCATION

- General 0515
- Administration 0514
- Adult and Continuing 0516
- Agricultural 0517
- Art 0273
- Bilingual and Multicultural 0282
- Business 0688
- Community College 0275
- Curriculum and Instruction 0727
- Early Childhood 0518
- Elementary 0524
- Finance 0277
- Guidance and Counseling 0519
- Health 0680
- Higher 0745
- History of 0520
- Home Economics 0278
- Industrial 0521
- Language and Literature 0279
- Mathematics 0280
- Music 0522
- Philosophy of 0998
- Physical 0523
- Psychology 0525
- Reading 0535
- Religious 0527
- Sciences 0714
- Secondary 0533
- Social Sciences 0534
- Sociology of 0340
- Special 0529
- Teacher Training 0530
- Technology 0710
- Tests and Measurements 0288
- Vocational 0747

LANGUAGE, LITERATURE AND LINGUISTICS

- Language
  - General 0679
  - Ancient 0289
  - Linguistics 0290
  - Modern 0291
- Literature
  - General 0401
  - Classical 0294
  - Comparative 0295
  - Medieval 0297
  - Modern 0298
  - African 0316
  - American 0591
  - Asian 0305
  - Canadian (English) 0352
  - Canadian (French) 0355
  - English 0593
  - Germanic 0311
  - Latin American 0312
  - Middle Eastern 0315
  - Romance 0313
  - Slavic and East European 0314

PHILOSOPHY, RELIGION AND THEOLOGY

- Philosophy 0422
- Religion
  - General 0318
  - Biblical Studies 0321
  - Clergy 0319
  - History of 0320
  - Philosophy of 0322
- Theology 0469

SOCIAL SCIENCES

- American Studies 0323
- Anthropology
  - Archaeology 0324
  - Cultural 0326
  - Physical 0327
- Business Administration
  - General 0310
  - Accounting 0272
  - Banking 0770
  - Management 0454
  - Marketing 0338
- Canadian Studies 0385
- Economics
  - General 0501
  - Agricultural 0503
  - Commerce-Business 0505
  - Finance 0508
  - History 0509
  - Labor 0510
  - Theory 0511
- Folklore 0358
- Geography 0366
- Gerontology 0351
- History
  - General 0578
  - Ancient 0579
  - Medieval 0581
  - Modern 0582
  - Black 0328
  - African 0331
  - Asia, Australia and Oceania 0332
  - Canadian 0334
  - European 0335
  - Latin American 0336
  - Middle Eastern 0333
  - United States 0337
- History of Science 0585
- Law 0398
- Political Science
  - General 0615
  - International Law and Relations 0616
  - Public Administration 0617
- Recreation 0814
- Social Work 0452
- Sociology
  - General 0626
  - Criminology and Penology 0627
  - Demography 0938
  - Ethnic and Racial Studies 0631
  - Individual and Family Studies 0628
  - Industrial and Labor Relations 0629
  - Public and Social Welfare 0630
  - Social Structure and Development 0700
  - Theory and Methods 0344
- Transportation 0709
- Urban and Regional Planning 0999
- Women's Studies 0453

THE SCIENCES AND ENGINEERING

BIOLOGICAL SCIENCES

- Agriculture
  - General 0473
  - Agronomy 0285
  - Animal Culture and Nutrition 0475
  - Animal Pathology 0476
  - Food Science and Technology 0359
  - Forestry and Wildlife 0478
  - Plant Culture 0479
  - Plant Pathology 0480
  - Plant Physiology 0817
  - Range Management 0777
  - Wood Technology 0746
- Biology
  - General 0306
  - Anatomy 0287
  - Biostatistics 0308
  - Botany 0309
  - Cell 0379
  - Ecology 0329
  - Entomology 0353
  - Genetics 0369
  - Limnology 0793
  - Microbiology 0410
  - Molecular 0307
  - Neuroscience 0317
  - Oceanography 0416
  - Physiology 0433
  - Radiation 0821
  - Veterinary Science 0778
  - Zoology 0472
- Biophysics
  - General 0786
  - Medical 0760

EARTH SCIENCES

- Biogeochemistry 0425
- Geochemistry 0996
- Geodesy 0370
- Geology 0372
- Geophysics 0373
- Hydrology 0388
- Mineralogy 0411
- Paleobotany 0345
- Paleoecology 0426
- Paleontology 0418
- Paleozoology 0985
- Palynology 0427
- Physical Geography 0368
- Physical Oceanography 0415

HEALTH AND ENVIRONMENTAL SCIENCES

- Environmental Sciences 0768
- Health Sciences
  - General 0566
  - Audiology 0300
  - Chemotherapy 0992
  - Dentistry 0567
  - Education 0350
  - Hospital Management 0769
  - Human Development 0758
  - Immunology 0982
  - Medicine and Surgery 0564
  - Mental Health 0347
  - Nursing 0569
  - Nutrition 0570
  - Obstetrics and Gynecology 0380
  - Occupational Health and Therapy 0354
  - Ophthalmology 0381
  - Pathology 0571
  - Pharmacology 0419
  - Pharmacy 0572
  - Physical Therapy 0382
  - Public Health 0573
  - Radiology 0574
  - Recreation 0575
  - Speech Pathology 0460
  - Toxicology 0383
- Home Economics 0386

PHYSICAL SCIENCES

Pure Sciences

- Chemistry
  - General 0485
  - Agricultural 0749
  - Analytical 0486
  - Biochemistry 0487
  - Inorganic 0488
  - Nuclear 0738
  - Organic 0490
  - Pharmaceutical 0491
  - Physical 0494
  - Polymer 0495
  - Radiation 0754
- Mathematics 0405
- Physics
  - General 0605
  - Acoustics 0986
  - Astronomy and Astrophysics 0606
  - Atmospheric Science 0608
  - Atomic 0748
  - Electronics and Electricity 0607
  - Elementary Particles and High Energy 0798
  - Fluid and Plasma 0759
  - Molecular 0609
  - Nuclear 0610
  - Optics 0752
  - Radiation 0756
  - Solid State 0611
- Statistics 0463

Applied Sciences

- Applied Mechanics 0346
- Computer Science 0984
- Engineering
  - General 0537
  - Aerospace 0538
  - Agricultural 0539
  - Automotive 0540
  - Biomedical 0541
  - Chemical 0542
  - Civil 0543
  - Electronics and Electrical 0544
  - Heat and Thermodynamics 0348
  - Hydraulic 0545
  - Industrial 0546
  - Marine 0547
  - Materials Science 0794
  - Mechanical 0548
  - Metallurgy 0743
  - Mining 0551
  - Nuclear 0552
  - Packaging 0549
  - Petroleum 0765
  - Sanitary and Municipal 0554
  - System Science 0790
- Geotechnology 0428
- Operations Research 0796
- Plastics Technology 0795
- Textile Technology 0994

PSYCHOLOGY

- General 0621
- Behavioral 0384
- Clinical 0622
- Developmental 0620
- Experimental 0623
- Industrial 0624
- Personality 0625
- Physiological 0989
- Psychobiology 0349
- Psychometrics 0632
- Social 0451

UNIVERSITY OF ALBERTA
RELEASE FORM

NAME OF AUTHOR: Emmanuel Benhin

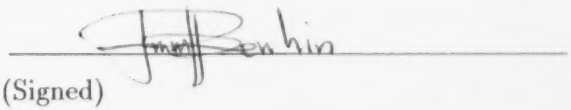
TITLE OF THESIS: **Application of Empirical Likelihood Estimation in Survival Data Analysis and Survey Sampling**

DEGREE FOR WHICH THESIS WAS PRESENTED: Master of Science

YEAR THE DEGREE GRANTED: 1994

Permission is hereby granted to THE UNIVERSITY OF ALBERTA LIBRARY to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only.

The author reserves other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without the author's writen permission.



(Signed)

Permanent Address:
P.O.Box 15079
Accra-North
Ghana

Date: August 30th, 1994

University of Alberta

Faculty of Graduate Studies and Research

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled **Application of Empirical Likelihood Estimation in Survival Data Analysis and Survey Sampling** submitted by **Emmanuel Benhin** in partial fulfillment of the requirements for the degree of **Master of Science**.

(Supervisor) Prasad N. G. N.

Karunamuni R.

Talwar P. P.

Date: Aug. 25, 1994

Abstract

This thesis has two parts: (i) Empirical likelihood inference in survival analysis, and (ii) Empirical likelihood estimation in survey sampling. In part (i), we obtain empirical likelihood inference on the median of lifetime distribution which is subject to right censoring at some specified time T_o. Simulation results presented in this report show that the empirical likelihood estimates perform well among a number of nonparametric competitors.

In part (ii), we use empirical likelihood method of estimation to estimate finite population characteristics from a sample survey data. In particular, we consider two-phase sampling design and we propose two alternative estimators in the presence of information on two auxiliary variables. The suggested empirical likelihood estimators are found to be more efficient than the ratio-type and the regression-type estimators suggested by Kiregyera (1980, 1984) and the ratio-cum-regression and regression-cum-regression estimators suggested by Sahoo et al. (1994).

ACKNOWLEDGMENTS

I wish to express my profound appreciation to Dr. N. G. N. Prasad and Dr. A. Wong for their invaluable guidance and considerable time spent in assisting me in preparing this thesis.

My special thanks go to my Heavenly Father who gave me the strength, good health, and peaceful environment for my work. Thanks also to my family for their wonderful faith in me and moreso their prayerful support.

Finally, I wish to express my gratitude to Ms. Rachael Ingrid Nartey for her enormous encouragement.

Contents

1 Introduction **1**

1.1 Censoring in Lifetime data 3

1.2 Models for Lifetime data 5

1.3 Basic Concepts of Lifetime Distribution 6

2 Statistical Inference **10**

2.1 Statistical Inference Procedures for Some Parametric Distributions . . 10

2.2 Nonparametric Models 15

2.3 Empirical Likelihood 19

3 Statistical Inference for the Median **22**

3.1 Empirical Likelihood Inference on the Median when $S(T_o)$ is known . 22

3.2 Empirical Likelihood Inference on the Median when $S(T_o)$ is unknown 26

3.3 Interval Estimation for the Median 27

3.4 Simulation Study 29

4 Estimation in Survey Sampling in the Presence of Auxiliary Information **34**

4.1 Ratio Method of Estimation 34

4.2 Regression Method of Estimation 36

4.3 Difference Method of Estimation 37

4.4 Empirical Likelihood Method of Estimation 37

4.5 Double Sampling 39

4.6 Estimation in Two-Phase Sampling using Two Auxiliary Variables . . 40

5 Empirical Likelihood Estimation in Two-Phase Sampling Using Two Auxiliary Variables **45**

5.1 Regression-thru-ratio Method of Estimation 45

5.2 Empirical Likelihood Regression-Thru-Ratio Method of Estimation . 47

5.3 Empirical Likelihood Regression-Thru-Regression Estimation 50
5.4 Numerical Illustrations 53
5.5 Concluding Remarks and Further Research 55
Bibliography 56

Chapter 1

Introduction

Empirical likelihood provides a way to find efficient estimators in many areas of statistical application when supplementary information is available. Recently, several researchers have studied how to use supplementary information to obtain better statistical inference. For example, Owen (1991) has considered a constrained empirical likelihood, such that if we knew the value of some functional of the distribution, it would be natural to consider only those distributions which are equal in distribution to the functional. This information should allow us to sharpen our inferences for other functions. The result is first order equivalent to conditioning on a sample value of the known functional. Haberman (1984) considers distributions that minimize Kullback-Leibler distance measure from an empirical distribution subject to linear constraints. Sheehy (1988) shows that the resulting distribution function estimate is asymptotically efficient. Qin (1991) shows that the constrained empirical likelihood distribution function is asymptotically equivalent to Haberman's estimate. Hence the constrained empirical likelihood estimate of distribution is efficient. In the context of sampling from a finite population, Kuk and Mak (1989) discuss inference on the population median in the presence of auxiliary information. Chen and Qin (1993) discussed the use of empirical likelihood estimation for finite populations and the effective usage of auxiliary information.

This thesis has two parts. The first part (Chapters 1 to 3), discusses the role of empirical likelihood in survival data analysis, in particular how it may be used in providing inference for the median or any other quantile of a lifetime distribution which may be subject to right censoring at some specified time T_o.

We must mention that the median survival time is one of the most useful and frequently reported characteristics from analysing survival or lifetime data. In the biomedical sciences, one of the major areas where survival data analysis is encountered, the estimate of the median survival time is often used for evaluation of the efficacy of a treatment for a chronic disease. For censored data this measure is cer-

tainly preferable to the estimate of the mean survival time because it is easy to obtain and has a clearer interpretation.

Even though it has become a common practice in the medical literature to give point estimates for the median survival time, biostatisticians (for example Peto *et al.*, 1977) have pointed out that point estimates of the median survival time can be seriously misleading because of their relatively large variation, particularly if the survival curve does not change rapidly near the median. It would be more appropriate therefore to provide interval estimates for the median.

Lifetime distributions may be represented by specific parametric models or by nonparametric models or by semiparametric models. In situations where parametric models can be used, inferences based on the models can be made easily at least in principle. Nonparametric models are employed when it is not feasible to use parametric models. Situations arise however, whereby lifetime distributions can be specified incompletely by a parametric model. The model obtained in such situations is what we refer to as the semiparametric model.

Chapter 1 discusses some basic censoring mechanisms and introduces some basic concepts of lifetime distribution. Chapter 2 deals mainly with statistical inference of the median using parametric models or nonparametric models. In Chapter 3 we turn our attention to a situation where a constrained empirical likelihood is employed for constructing interval estimates or obtaining hypothesis tests for the median survival time when we have a right censored data, we also discuss the simulation study conducted using the various methods mentioned in this report.

The results presented in Chapter 3 indicate that the empirical likelihood method provides very good coverage probabilities at almost all levels of censoring which are less than fifty percent for testing the median survival time. It is also observed that the corresponding average interval lengths for the confidence intervals are shorter in most cases than their counterparts using the nonparametric methods. It is interesting to note that not only is the method applicable to the median but also to other quantiles provided these quantiles are known to be in the uncensored part of the lifetime distribution.

The second part of this thesis (Chapters 4 to 5), discusses the role of empirical likelihood in estimation in survey sampling, in particular when *two-phase sampling* is

used in the presence of two auxiliary variables to estimate population characteristics.

In Chapter 4, we review conventional methods of estimation namely, ratio method of estimation, difference method of estimation and regression method of estimation. We also introduce empirical likelihood method of estimation using auxiliary information. All these methods require the knowledge of the population mean, $\bar{X}$, of the auxiliary variate x. However, this information sometimes, is not available and in situations like this, it is often convenient and cheaper to take a large preliminary sample in which only x is measured. The purpose of this is to furnish a good estimate of $\bar{X}$. In a survey, whose objective is to make estimates for the study variate y, it may pay to devote part of the resources to this preliminary sample, although this means that the sample in the main survey on y must be decreased. This technique is known as *double sampling* or *two-phase sampling*.

Sometimes even if $\bar{X}$ is unknown, information on cheaply ascertainable variable z, closely related to x but compared to x remotely related to y, is available on all units of the population. Chand (1975), Kiregyera (1980, 1984), Prasad and Srivankataramana (1980), Srivastava et al. (1988, 1990) and Sahoo et al. (1994) consider this type of situation. The methods of estimation under this technique proposed by Chand, kiregyera and Sahoo et al. are also discussed in Chapter 4.

In Chapter 5, we apply empirical likelihood principle to obtain two alternative estimators under double sampling technique when information on auxiliary variables is available. We observe that the estimators that we propose are more efficient than those of Chand, Kiregyera and Sahoo et al.. Numerical illustrations using real data sets to compare the proposed estimators with the other conventional estimators are presented.

The next few sections of this chapter are devoted to introduce basic concepts in survival data analysis.

1.1 Censoring in Lifetime data

In statistical investigations one encounters various kinds of data, one kind of which is the "lifetime" data which are also referred to as "survival time" or "failure time" data. All these names refer to similar kind of data from different application areas such as engineering, medical sciences, biological sciences and so on. Lifetime data

are obtained from experiments where one is interested in the time of occurrences of some event of interest for some population of individuals. These times of occurrences of these events are termed "lifetimes" and the data collected from such experiments are referred to as lifetime data. Lifetime data often come with a feature that creates special problems in their analysis. This feature is known as censoring. It occurs when exact lifetimes are known for only a portion of the individuals under study or investigation; the remainder of the lifetimes are known only to exceed or fall below certain values. An observation may be right or left censored. It is said to be right censored, say at L if the exact value of the observation is not known but only that it is known to be greater than or equal to L. Similarly, an observation is left censored at L if it is known only that it is less than or equal to L. For our purpose we shall consider only the right censoring feature. There are several types of censoring and we shall briefly discuss only a few, the Type I censoring and the Type II censoring.

Type I Censoring

Type I (or "time") censored sample is obtained when experiments are run over a fixed time period in such a way that an individual's lifetime will be known exactly only if it is less than some predetermined value.

For instance, in a life test experiment n items may be placed on test, but a decision made to terminate the test after a time L has elapsed. Lifetimes will then be known exactly only for those items that fail by time L. This type of censoring frequently arises in medical research where, for example, a decision is made to terminate a study at a date on which not all the individuals' lifetimes will be known.

In general, each item in the experiment may have its own specific censoring time L_i since all items may not start on the test on the same date. Stated formally, a Type I censored sample is one that arises when items $1, \ldots, n$ are subjected to limited periods of observation $L_1, \ldots, L_n$, so that an items lifetime T_i is observed only if $T_i \leq L_i$. When all of the L_i's are equal, we sometimes say that the data are singly Type I censored, to distinguish this from the general case. It should be noted that with Type I censoring the number of exact lifetimes observed is random in contrast to the case of Type II censoring where it is fixed.

Type II Censoring

A Type II censored sample is one for which only r smallest observations in a sample of n items are observed ($1 \leq r \leq n$). Type II censoring is often used in experiments involving life testing; a total of n items is placed on test, but instead of continuing until all n items have failed, the test is terminated at the time of the rth failure. Such tests can save time and money, since it could take a very long time for all items to fail in some instances. It will be seen that the statistical treatment of Type II censored data is at least in principle, straightforward.

There are other types of censoring that are generalizations of either the Type I or the Type II censoring. The progressive Type II as the name implies is a generalization of the Type II censoring and the random censoring, which is a conditional form of the Type I censoring.

1.2 Models for Lifetime data

The two kinds of models generally encountered in analysis of lifetime data are the parametric family of models and the nonparametric models. These methods are used as far as the data suggest their feasibility and appropriateness. We would discuss each of these models.

Parametric Models

Several parametric models are commonly used in the analysis of lifetime data. Among the univariate models a few occupy a central role because of their demonstrated usefulness in a wide range of situations. Foremost in this category are the exponential, Weibull, gamma and log-normal distributions. Motivation for using a particular model in a given situation is often mainly empirical and sometimes their simplicity when the model has been found to satisfactorily describe the distribution of the lifetimes in the population under study. In fact this means that the choice of a particular model does not imply the authenticity or absolute correctness of the model. There is however, a theoretical motivation also. References to this can be found in Johnson and Kotz (1970), which extensively catalogs mathematical and statistical properties of most well-known distributions.

Nonparametric Models

Among the most commonly used and reported estimate in analysing lifetime data whose distribution is assumed to be a nonparametric model is the "product- limit" estimate of the survivor function. This estimate is normally referred to as the Kaplan-Meier estimate, named after the authors who first discussed its properties. This estimate forms the basis for most nonparametric models used in statistical inference. A few of such models are by Emerson (1982), Breslow and Crowley (1974) and Brookmeyer and Crowley (1982).

1.3 Basic Concepts of Lifetime Distributions

Continuous Models

Consider a case of a single lifetime variable T, a nonnegative random variable representing the lifetimes of individuals in some population. We begin by assuming that T is continuous. Unless otherwise stated all are defined over the interval $[0, \infty)$. Let $f(t)$ denote the probability density function (p.d.f.) of T and let the distribution function be

$$F(t) = P(T \leq t) = \int_0^t f(x)dx.$$

Definition I

The *survival function* $S(t)$ is defined as the probability that a subject's lifetime is at least t units;

$$S(t) = P(T \geq t) = \int_t^\infty f(x)dx.$$

Some Properties of S(t)

1. $0 \leq S(t) \leq 1$
2. $S(0) = 1$
3. $S(\infty) = \lim_{t \to \infty} S(t) = 0$
4. S(t) is a monotone non-increasing function.

The *p-th* quantile of the distribution of T is the value t_p such that $S(t_p) = 1 - p; 0 \leq p \leq 1$.

Discrete Models

Situation arises where lifetimes are grouped or when lifetime refers to an integral number of cycles of some sort, then it may be desired to treat T as a discrete random

variable. Suppose T can take on values $t_1, t_2, \ldots$, with $0 \leq t_1 \leq t_2 \leq \cdots$, and let the probability function (p.f.) be

$$P_j = p(t_j) = P(T = t_j) \qquad j = 1, 2 \ldots .$$

The survivor function is then given by

$$S(t) = P(T \geq t) = \sum_{j; t_j \geq t} p(t_j).$$

$S(t)$ for the discrete random variable T has all the first three properties of the survival function of the continuous distribution and in addition it is a monotone non-increasing left-continuous function.

Definition II

Suppose $x_1, \ldots, x_n$ are random sample from a distribution with a p.d.f. or probability function (p.f.) $f(x; \theta)$ where θ is a parameter taking on values in a set Ω. The likelihood for θ is defined as

$$L(\theta) = c \prod_{i=1}^{n} f(x_i; \theta),$$

where $c > 0$ is any arbitrary multiplicative constant. Without loss of generality, c is taken to be 1 hereafter. We note that $L(\theta) \geq 0$ and it is a function of θ which depends on $x_1, \ldots, x_n$.

We now present likelihood arguments for Type I and Type II censored data. Details for these may be found in Lawless (1982). We first consider the Type II censored case. It must be stressed that with the Type II censoring, the number of observations r is decided before the data are collected. Formally, the data consist of r smallest lifetimes $T_{(1)} \leq T_{(2)} \leq \cdots \leq T_{(r)}$ out of a random sample of n lifetimes $T_1, \ldots, T_n$ from a given lifetime distribution. If $T_1, \ldots, T_n$ are independently and identically distributed and have a continuous distribution with p.d.f. $f(t)$ and survivor function $S(t)$, it follows from general results on order statistics that the joint p.d.f. of $T_{(1)}, \ldots, T_{(r)}$ is

$$f(t_{(1)}) \cdots f(t_{(r)}) \left[S(t_{(r)})\right]^{n-r}.$$

It is worth while to mention that for any given lifetime data which is Type II censored and whose distribution may be considered to be continuous and of some

kind of parametric nature, the above definition may be employed in deriving sampling properties of various statistics and statistical inference may be conducted for the distribution.

In dealing with Type I censoring, suppose that there are n individuals under study and that associated with the ith individual is a lifetime T_i and a fixed censoring time L_i. The T_i's are assumed to be i.i.d. with p.d.f. $f(t)$ and survivor function $S(t)$. The exact lifetime T_i of an individual will be observed only if $T_i \leq L_i$. The data from such a setup can be conveniently represented by the n pairs of random variables (t_i, δ_i), where

$$t_i = \min(T_i, L_i) \quad \text{and} \quad \delta_i = \begin{cases} 1 & \text{if } T_i \leq L_i \\ 0 & \text{if } T_i > L_i. \end{cases}$$

That is, δ_i indicates whether the lifetime T_i is censored or not, and t_i is equal to T_i if it is observed, and to L_i if it is not. the joint p.d.f. of t_i and δ_i is

$$f(t_i)^{\delta_i} S(L_i)^{1-\delta_i}.$$

To see this, note that t_i is a mixed random variable with a continuous and a discrete component. For the discrete part we have

$$\begin{aligned} P(t_i = L_i = i) &= P(\delta_i = 0) \\ &= P(T_i > L_i) \\ &= S(L_i). \end{aligned}$$

From values $t_i < L_i$ the continuous p.d.f. is

$$\begin{aligned} P(t_i|\delta_i = 1) &= P(t_i|t_i < L_i) \\ &= \frac{f(t_i)}{1 - S(t_i)}, \end{aligned}$$

where for convenience we have used the notation $P(t_i|\delta_i = 1)$ to represent the probability density function of t_i, given that $t_i < L_i$. The distribution of (t_i, δ_i) thus has components:

$$P(t_i = L_i, \delta_i = 0) = P(\delta_i = 0) = S(L_i) \text{ and}$$
$$P(t_i, \delta_i = 1) = P(t_i|\delta_i = 1)P(\delta_i = 1) = f(t_i), \text{ for } t_i < L_i.$$

These expressions can be combined into the single expression

$$P(t_i, \delta_i) = f(t_i)^{\delta_i} S(L_i)^{1-\delta_i},$$

and if pairs (t_i, δ_i) are independent, then the likelihood function is

$$\prod_{i=1}^{n} f(t_i)^{\delta_i} S(L_i)^{1-\delta_i}.$$

This forms the basis of inference for lifetime data which is Type I censored and known to have a parametric distribution with p.d.f. $f(t)$.

Chapter 2

Statistical Inference

2.1 Statistical Inference Procedures for some Parametric Distributions

Single samples of noncensored data or of Type II censored data will be the focus of our discussion. We shall concentrate mostly on the exponential model which is one of the commonly used distributions. The methods described here can be extended, with a few modifications to either the Weibull, gamma, or the log-normal distributions.

Exponential Model

The probability density function of the exponential model with mean θ is given by

$$f(t;\theta) = \theta^{-1} \exp(-t/\theta) \qquad t \geq 0,\ \theta > 0. \tag{2.1}$$

With noncensored data samples inference procedures are simple and well known. Let $T_1, \ldots, T_n$ be a random sample from the model (2.1), then, by Definition II, the likelihood function for this sample is

$$L(\theta) = \prod_{i=1}^{n} f(t_i;\theta) = \frac{1}{\theta^n} \exp\left(-\sum_{i=1}^{n} \frac{t_i}{\theta}\right). \tag{2.2}$$

The maximum likelihood esimator (m.l.e.) of θ, obtained by maximizing (2.2), is easily found to be $\hat{\theta} = t/n$, where $T = \sum_{i=1}^{n} T_i$ is the minimal sufficient statistic and $t = \sum_{i=1}^{n} t_i$ is the minimal sufficient statistic obtained from the sample. Since the $T_i/\theta\,(i = 1, \ldots, n)$ are independent standard exponential variables, T/θ has a one-parameter gamma distribution with index parameter n. Equivalently, $2T/\theta$ is distributed as a chi-square distribution with $2n$ degrees of freedom.

It can easily be shown that similar results hold for Type II censored sampling. Suppose that only the first r observations $T_{(1)} \leq T_{(2)} \leq \cdots \leq T_{(r)}$ are available in a total sample of size n. Using the arguments of the likelihood for the Type II censored

data as described in the previous Chapter, we obtain the likelihood for the above sample to be

$$\begin{aligned} & \left[\prod_{i=1}^{r} \frac{1}{\theta} \exp\left(-\frac{t_{(i)}}{\theta}\right)\right] \left[\exp\left(-\frac{t_{(r)}}{\theta}\right)\right]^{n-r} \\ = \quad & \frac{1}{\theta^r} \exp\left[-\left(\sum_{i=1}^{r} t_{(i)} + (n-r)t_{(r)}\right)/\theta\right]. \end{aligned} \tag{2.3}$$

Let

$$T = \sum_{i=1}^{r} T_{(i)} + (n-r)T_{(r)},$$

and (2.3) can be rewritten as

$$L(\theta) = \frac{1}{\theta^r} e^{-\frac{t}{\theta}}. \tag{2.4}$$

Clearly the m.l.e. for θ is $\hat{\theta} = t/r$. It follows that

$$2T/\theta \sim \chi^2_{(2r)},$$

which can be used to obtain inference for the exponential distribution with Type II censored data.

Tests and Confidence Intervals

Tests and confidence intervals for θ, the p-th quantile, $t_p = \theta[-\log(1-p)]$ and the survival function at time t_0, $S(t_0) = \exp(-t_0/\theta)$ can be obtained using the pivotal quantity $2T/\theta$.

Suppose $[A(T), B(T)]$ is the $(1-\alpha)*100\%$ confidence interval for θ. Then:

1. A $(1-\alpha)*100\%$ confidence interval for the p-th quantile is

$$[(-\log(1-p))A(T), (-\log(1-p))B(T)].$$

2. A $(1-\alpha)*100\%$ confidence interval for $S(t_0)$ is

$$[\exp(-t_0/A(T)), \exp(-t_0/B(T))].$$

Location-Scale Parametric Models

Some of the location-scale parametric models employed in survival data analysis are the the normal distribution, the log-normal distribution, Weibull and the extreme value distributions and so on.

A family of distributions with location parameter μ $(-\infty < \mu < \infty)$ and scale parameter θ $(\theta > 0)$, has a p.d.f. of the form

$$f(x;\mu,\theta) = \frac{1}{\theta} g\left(\frac{x-\mu}{\theta}\right) \qquad -\infty < x < \infty \tag{2.5}$$

and survivor function $G[(x-\mu)/\theta]$, where

$$G(y) = \int_y^\infty g(z)dz.$$

Equivariant Estimators

Suppose that $x_1 \leq x_2 \leq \ldots \leq x_r$ is a Type II censored sample consisting of r smallest observations in a total sample of size n $(r \leq n)$ from a distribution with p.d.f. given by (2.5). Let $\tilde{\mu} = \tilde{\mu}(x_1, \ldots, x_r)$ and $\tilde{\theta} = \tilde{\theta}(x_1, \ldots, x_r)$ be estimators of μ and θ possessing the following invariance properties;

$$\tilde{\mu}(bx_1 + c, \ldots, bx_r + c) = b\tilde{\mu}(x_1, \ldots, x_r) + c \tag{2.6}$$

$$\tilde{\theta}(bx_1 + c, \ldots, bx_r + c) = b\tilde{\theta}(x_1, \ldots, x_r) \tag{2.7}$$

for any real constants c $(-\infty < c < \infty)$ and b $(b > 0)$. Such estimators are termed "equivariant". The requirements (2.6) and (2.7) are simply that location and scale changes on the data should induce the same location change in $\tilde{\mu}$ and the same scale change in $\tilde{\theta}$. The following theorem helps us to construct interval estimates or obtain test statistics for the location and scale parameters and also for the *p-th* quantile.

Theorem 2.1.1 *If $\tilde{\mu}$ and $\tilde{\theta}$ are equivariant of μ and θ, based on a Type II censored sample $t_1 \leq \cdots \leq t_r$ from (2.5), then*

1. *$z_1 = \tilde{\mu}/\tilde{\theta}$, $z_2 = \tilde{\theta}/\theta$ and $z_3 = (\tilde{\mu} - \mu)/\theta$ are pivotal (parameter free) quantities.*
2. *The quantities $a_i = (t_i - \tilde{\mu})/\tilde{\theta}$ form a set of ancillary statistics (i.e statistics whose distribution does not depend on μ or θ), of which only $r-2$ are functionally independent.*

For the proof of Theorem 2.1.1, see Statistical Models and Methods for Lifetime Data by Lawless (1982).

Using the pivotals z_1 and z_2, based on a particular form of equivariant estimators, and from the above theorem, one can construct confidence intervals for μ and θ. Similarly for the *p-th* quantile for the model (2.5) $t_p = \mu + w_p\theta$, where w_p satisfies $G(w_p) = 1-p$, one can construct confidence intervals for t_p using the pivotal quantity

$$\begin{aligned} z_p &= \frac{(\tilde{\mu} - \mu) - w_p\theta}{\tilde{\theta}} \\ &= \frac{\tilde{\mu} - t_p}{\tilde{\theta}}, \end{aligned} \tag{2.8}$$

where z_p follows noting that $z_p = z_1 - w_p z_2^{-1}$.

Although confidence intervals for μ, θ or t_p, can, in principle be obtained from the pivotals z_1, z_2 and z_p, a practical difficulty is that their distributions are mathematically intractable. There are other methods with stringent assumptions which are used to construct confidence intervals for μ, θ or t_p. Some of these procedures or methods are discussed below.

Conditional Method

Let $\tilde{\mu}$ and $\tilde{\theta}$ be equivariant estimators of μ and θ under the conditions of *Theorem 2.1.1.* Then the joint p.d.f. of $z_1, z_2, a_1, \ldots, a_{r-2}$ is of the form

$$k(\mathbf{a}, r, n) z_2^{r-1} \left(\prod_{i=1}^{r} g(a_i z_2 + z_1 z_2) \right) [G(a_r z_2 + z_1 z_2)]^{n-r},$$

where $k(\mathbf{a}, r, n)$ is a function of $a_1, \ldots, a_{r-2}, r$, and n only. The conditional p.d.f. of (z_1, z_2) given $\mathbf{a} = (a_1, \ldots, a_r)$ is also of the form as given above. Except for the case of uncensored samples from the normal distribution, it turns out that the manner in which the a_i's enter $k(\mathbf{a}, r, n)$ and the form of the rest of the above p.d.f. make it impossible to integrate out $a_1, \ldots, a_{r-2}$, so one cannot obtain the distributions of z_1 and z_2 analytically.

It must be mentioned that confidence intervals for μ, θ or t_p obtained from the conditional distributions of z_1, z_2, or z_p, given $\mathbf{a}$ are much easier to calculate than using the unconditional method as mentioned above. Detail discussions of this method can be found in Lawless (1982)

Approximation to Distribution of Pivotals

As mentioned earlier, obtaining exact distributions of z_1, z_2 and z_p are either not feasible analytically or even when feasible are complex to handle computationally. So that for all cases of practical importance, approximation to these distributions are considered. There are several of such approximations. In fact in most situations little is lost by using the approximate procedures, as long as the sample size is sufficiently large. This is to ensure that the approximations give the required accuracy. Three of the widely used large-sample methods will be discussed shortly.

Asymptotic and Large-sample Methods

Let $\hat{\theta}$ be a point in Ω at which $L(\theta)$ is maximized; $\hat{\theta}$ is called a maximum likelihood estimate (m.l.e.) of θ. Under regularity conditions $\hat{\theta}$ exists and uniquely defined. It is often convenient to work with $\log L(\theta)$ which is also maximized at $\hat{\theta}$, and in many cases $\hat{\theta}$ can be readily found by solving the so-called maximum likelihood equations $U(\theta) = 0$, where

$$U(\theta) = \frac{d \log L(\theta)}{d\theta}.$$

$U(\theta)$ has mean 0 and variance $I(\theta) = E\left(-\frac{d^2 \log L(\theta)}{d\theta^2}\right)$; $I(\theta)$ is called the Fisher Information. Under mild regularity conditions $\hat{\theta}$ is a consistent estimator of θ. In addition, several other asymptotic results hold that lead to useful inference procedures.

First of all, $U(\hat{\theta})$ is asymptotically $N[0, I(\theta)]$. This means that under the hypothesis $H_0 : \theta = \theta_0$

$$\frac{U^2(\theta_o)}{I(\theta_o)}$$

is asymptotically $\chi^2_{(1)}$. This can be used to test H_o and to obtain confidence intervals for θ. This procedure is called the Score procedure (also known as Rao's statistic).

Secondly, tests and estimates can also be based on $\hat{\theta}$. Under mild regularity conditions $\hat{\theta}$ is asymptotically $N[\theta, I^{-1}(\theta)]$. Thus under $H_o : \theta = \theta_o$

$$(\hat{\theta} - \theta_o)^2 I(\theta_o)$$

is asymptotically $\chi^2_{(1)}$. The procedure is attributed to Wald. It can also be used to test H_o and to obtain confidence intervals for θ.

We must mention that the Fisher's information, $I(\theta)$, is usually hard to calculate

so an observed information is normally used. This is given by

$$j(\hat{\theta}) = \frac{d^2 \log L}{d\theta}|_{\hat{\theta}}.$$

A third procedure for testing and providing interval estimates for θ is the likelihood ratio method. It can be shown that under $H_o : \theta = \theta_o$

$$\Lambda = -2 \log \left(\frac{L(\theta_o)}{L(\hat{\theta})} \right)$$

is asymptotically $\chi^2_{(1)}$. This is attributed to Wilks (1938), who provided the general theorem and the proof of this procedure.

The adequacy of the procedures mentioned above in finite samples changes from model to model and so it is difficult to make general statements. However, the distributions of the likelihood ratio statistic often appear to approach their limiting distribution considerably faster than the distribution of $\hat{\theta}$ and $U(\hat{\theta})$ to their limiting distributions. It is also important to note that the third method is not affected by reparameterization of the model, whereas the first and the second methods are.

Procedures based on the asymptotic normality of $\hat{\theta}$ are often preferred in view of their computational simplicity. Likelihood ratio and scores procedures for finding confidence intervals usually require more computation. However, the likelihood ratio method often produces intervals with closer to the nominal coverage probabilities because of the superior approximations that are provided by its asymptotic distributions whereas the score method exhibits this property only on few occasions.

For location-scale model, DiCiccio, Field and Fraser (1990) provide some higher order approximations for tail probability. Numerical procedure for their method is given in Wong (1992). This higher order method is extremely accurate but may be more complicated to obtain. Also Bartlett correction can be used to adjust the likelihood ratio method but again this is hard to calculate.

Empirical likelihood ratio method is the nonparametric version of the parametric likelihood ratio method. We employ this procedure to derive our main results.

2.2 Nonparametric Models

When circumstances do not support the adoption of a fully parametric lifetime distribution model, one often turns to nonparametric or distribution-free methods.

Nonparametric Estimation of the Survival Function

Suppose there are no censored observations in a sample of size n, then the empirical survivor function (ESF) is defined as

$$\hat{S}(t) = \frac{\text{Number of observations} \geq t}{n} \qquad t \geq 0. \tag{2.9}$$

This is a step function that decreases by $1/n$ just after each observed lifetime if all observations are distinct. More generally, if there are d lifetimes equal to t, the ESF drops by d/n just past t.

When dealing with censored data, some modifications of the (2.9) is necessary, since the number of lifetimes greater than or equal to t will not generally be known exactly. The modification of (2.9) described here has come to be called the "product-limit" (PL) estimate of the survivor function or, sometimes, the Kaplan-Meier estimate, from the authors who first discussed its properties (Kaplan and Meier, 1958). We give a formal definition for this estimate.

Definition III

Suppose that there are observations on n items and that there are $k(k \leq n)$ distinct times $t_1 < t_2 < \cdots < t_k$ at which deaths occur. The possibility of there being more than one death at t_i is allowed, and we let d_i represent the number of deaths at t_i. In addition to the lifetimes $t_1, \ldots, t_k$, there are also censoring times L_i for items whose lifetimes are not observed. The Kaplan-Meier estimate of $S(t)$ is defined as

$$\hat{S}(t) = \prod_{j:t_j<t} \frac{n_j - d_j}{n_j}, \tag{2.10}$$

where n_j is the number of items at risk at t_j, that is, the number of items alive and uncensored just prior to t_j. For various types of censoring and under quite general conditions, the Kaplan-Meier estimate of the survivor function, possesses a number of large sample properties. Breslow and Crowley (1974) showed that for random censoring model and under some general conditions, the Kaplan-Meier estimate is consistent and normally distributed. Similar results were obtained for fixed censoring time model by Meier (1975). When there is extreme censoring on the right however, the Kaplan-Meier (or product-limit) estimator is known to be a biased estimator of the survival function. Several modifications of the Kaplan-Meier estimator have been examined and compared with respect to bias and mean square error in Moeschberger

and Klein (1985). We provide an alternative approach using empirical likelihood. This is discussed in detail in the next Chapter.

Nonparametric Interval Estimates of Distribution Quantiles

Uncensored Data

Distribution-free confidence intervals for the *p-th* quantile of a continuous lifetime distribution can be readily obtained when the data are complete or Type II censored. Two-sided intervals for t_p are, for example, of the form $[t_{(r)}, t_{(s)})$, where $1 \leq r < s \leq n$ and $t_{(1)} < \ldots, < t_{(n)}$ are the ordered observations in a random sample of size n from the distribution in question.

To determine the confidence coefficient for an interval of the form as indicated above, suppose X is the number of observations in a random sample of size n that is less than or equal to t_p; X has a binomial distribution with probability function $\binom{n}{x} p^x(1-p)^{n-x}$. The inequality $t_{(r)} \leq t_p \leq t_{(s)}$ is satisfied by the sample if and only if $r \leq X \leq s-1$, and hence

$$P[t_{(r)} \leq t_p \leq t_{(s)}] = \sum_{x=r}^{s-1} \binom{n}{x} p^x(1-p)^{n-x}. \tag{2.11}$$

By utilizing the well-known relationship between the binomial distribution and the incomplete beta function, we can write (2.11) in the alternate form

$$P[t_{(r)} \leq t_p \leq t_{(s)}] = B_p(r, n-r+1) - B_p(s, n-s+1), \tag{2.12}$$

where $B_p(n, x)$ is the incomplete beta function given by

$$B_p(x, n-x+1) = \sum_{j=x}^{n} \binom{n}{j} p^j(1-p)^{n-j},$$

where $1 \leq x \leq n$.

The interval for t_p is thus distribution-free and has confidence coefficient $(1-\alpha) * 100\%$ given by (2.11). Confidence intervals can be calculated directly or obtained from tables of the binomial distribution or incomplete beta function.

It must be noted that for a given p and n, it will be possible to find intervals for t_p only for certain values of α. For example, if a two-sided .90 confidence interval for

t_p is required then r, s and n can be selected to make α as close as possible to .90 though it may not be possible to make it exactly .90.

When the data are Type II censored, the method just described still apply, provided, of course, that the experiment continue until the necessary order statistics $t_{(r)}$ and $t_{(s)}$ in (2.11) are observed.

For the past few years, several papers have been published on the testing and construction of interval estimates for the median of survival data using nonparametric methods. A few of these were by Brookmeyer and Crowley (1982), Efron (1981), Emerson (1982), Reid (1981) and Simon and Lee (1982). Of interest to our discussion we review the methods proposed by Emerson (1982) and Brookmeyer and Crowley (1982).

Emerson's method generalizes the usual method of inverting the sign test in the construction of the confidence interval for the median, to allow for right censoring, by using the Kaplan-Meier estimate of the survival function to approximate the number of failures beyond the hypothesized median. The method is applicable for both discrete and continuous distributions. The following briefly illustrates the method.

First, assuming that all observations are complete, and consider test of H_o: Median T $= \theta$ versus H_1: Median T $\neq \theta$. The usual test $N_+ = \sum I(T_k > \theta)$, the number of observations greater than θ, T_k is the kth observed lifetime. If the T_k have a continuous distribution, N_+ is $Bin(n, \frac{1}{2})$ under H_o. That is,

$$P(N_+ = l) = \left(\begin{array}{c} n \\ l \end{array} \right) \left(\frac{1}{2}\right)^n.$$

Hence reject H_o when either $N_+ \leq a$ or $N_+ \geq b$, where a and b are determined from

$$P\left\{Bin(n, \frac{1}{2}) \leq a\right\} \simeq \frac{1}{2}\alpha \simeq P\left\{Bin(n, \frac{1}{2}) \geq b\right\}.$$

When T is discrete and $P_r(T = \theta) > 0$, a decision must be made as to how the failures at θ are handled. Let N_- be the number of observations less than θ, and let N_o be the number tied at θ. Then the statistic $N_* = max(N_-, N_+)$ will be used, which has $Bin(n, \frac{1}{2})$ under H_o. The hypothesis is rejected when N_* is sufficiently large. A consequence of the use of N_* is that endpoints of the intervals are always included in the confidence sets.

The above results are easily extended to censored observations when censoring occur only to the right of the hypothesized median. In this situation unknown failure times are greater than the median and the preceeding method apply implicitly. This method with some modifications extends to a general case where observations are censored before time θ. For a detail discussion on this, the reader may refer to Emerson (1982).

We now discuss the method proposed by Brookmeyer and Crowley (1982). Their method gives a nonparametric asymptotic confidence interval for the median survival time when the data are subject to right censoring. They showed that for testing the hypothesis H_o: Median of survival function $S^o = M$ versus H_1: Median of survival function $S^o \neq M$, the test statistic

$$T = \frac{\{\hat{S}^o(M) - S^o(M)\}^2}{\{\hat{S}^o(M)\}^2} \sum_{\substack{\text{distinct} \\ X_i \leq M}} \frac{d_i}{N_x(X_i)\{N_x(X_i) + d_i\}}$$

is approximately $\chi^2_{(1)}$. Under the null hypothesis $S^o(M) = 1/2$, and for an approximate α-level test, the procedure is not to reject H_o when

$$\left(\hat{S}^o(M) - \frac{1}{2}\right)^2 \leq C_\alpha \left(\hat{S}^o(M)\right)^2 \sum_{\substack{\text{distinct} \\ X_i \leq M}} \frac{d_i}{N_x(X_i)\{N_x(X_i) + d_i\}}$$

where

$\hat{S}^o(M)$ is the Kaplan-Meier estimate evaluated at M,

$\{X_i\}$ are the observed survival times,

d_i is the number of observed deaths at X_i and

$N_x(X_i)$ is the number of observed survival times larger than X_i.

The test statistic above may be used to construct the confidence interval for the median. Detail discussion for the method can be found in Brookmeyer and Crowley (1982).

2.3 Empirical Likelihood

Empirical likelihood, a nonparametric method for constructing confidence intervals and obtaining statistical tests was introduced by Owen (1988). Properties of em-

pirical likelihood are described in Owen (1990, 1991), Hall (1990), DiCiccio and Romano (1989) and DiCiccio, Hall and Romano (1989, 1991). The use of the empirical likelihood has been extended to semiparametric models by Qin and Lawless (1991). We will look at the definition and the main features of the empirical likelihood.

Definition and Main Features of Empirical Likelihood

Consider a random sample $x_1, x_2, \ldots, x_n$ of size n drawn from an unknown d-variate distribution F_o having mean μ_o and nonsingular covariance matrix Σ_o. Denote the j-th component as $x_i^j, (i = 1, \ldots, n)$, so that $x_i = (x_i^1, \ldots, x_i^d)^\tau$. Let L be the empirical likelihood function for the mean. For a specific vector $\mu = (\mu^1, \ldots, \mu^d)^\tau$, $L(\mu)$ is defined to be the maximum value of $\prod p_i$ over all vectors $p = (p_1, \ldots, p_n)$ that satisfy the constraints

$$\sum_{i=1}^{n} p_i = 1, \quad \sum_{i=1}^{n} x_i p_i = \mu.$$

As noted by Owen (1988, 1990), a unique maximum exists, provided that μ is inside the convex hull of the points $x_1, \ldots, x_n$. That is, $x_{(1)} < \mu < x_{(n)}$. An explicit expression for $L(\mu)$ can be derived by a Lagrange multiplier argument. The maximum of $\prod_{i=1}^n p_i$ subject to the above constraints is attained when

$$p_i = p_i(\mu) = n^{-1}\{1 + t^\tau(x_i - \mu)\}^{-1} \qquad i = 1, \ldots, n \tag{2.13}$$

where $t = t(\mu)$ is a d-dimensional column vector which is the Lagrange multiplier satisfying

$$\sum_{i=1}^{n} \{1 + t^\tau(x_i - \mu)\}^{-1}(x_i - \mu) = 0. \tag{2.14}$$

Since $\prod_{i=1}^n p_i$ attains its largest values over all vectors $p = (p_1, \ldots, p_n)$ satisfying $\sum_{i=1}^n p_i = 1$. When $p_i = n^{-1}(i = 1, \ldots, n)$, it follows that the empirical likelihood function $L(\mu)$ is maximized at $\hat{\mu} = \bar{x} = n^{-1}\sum_{i=1}^n x_i$ and $L(\hat{\mu}) = n^{-n}$.

The empirical likelihood ratio at the point μ is

$$\frac{L(\mu)}{L(\hat{\mu})} = \prod_{i=1}^{n} \{1 + t^\tau(x_i - \mu)\}^{-1} \tag{2.15}$$

and minus twice the logarithm likelihood ratio is

$$W(\mu) = 2\sum_{i=1}^{n} \log\{1 + t^\tau(x_i - \mu)\}. \tag{2.16}$$

Under the appropriate regularity conditions, Owen (1988, 1991) has proved that a version of Wilk's theorem for nonparametric models holds, that is to say, under H_o: $\mu = \mu_o$, $W(\mu_o) \rightarrow \chi^2_d$.

The idea of empirical likelihood has been extended to allow for a more general parametric constraints by Qin and Lawless (1991).

Chapter 3

Statistical Inference for the Median

Main Results

In survival data analysis, we sometimes encounter situations whereby the distribution of lifetime data may be subject to right censoring at a fixed or specified time T_o such that all the censored observations are beyond T_o. In situations like these, the survival function at T_o may or may not be known. For the purposes of our study, we consider lifetime data that are of a counting type or may be made so by grouping.

In this Chapter we propose empirical likelihood inference on the median of a lifetime distribution of the kind as described above. We would discuss this method under two cases: Case I, when the survival function at T_o is known and Case II, when the survival function at T_o is unknown.

Even though the method is discussed fully for cases of large samples, it must be emphasized that it is also very much reliable for data of intermediate and small sample sizes. It is also important to note that, these results can be applied to any other quantiles as long as the quantiles of interest are known to occur in the noncensored part of the distribution.

The next section gives a detail discussion on Case I and Section 3.2 extends the discussion to Case II.

3.1 Empirical Likelihood Inference on the Median when $S(T_o)$ is known

Let $t_1, \ldots, t_r$ denote the distinct times of observed failures (which are discrete or may be made so by grouping) before the fixed time T_o with probabilities $p_1, \ldots, p_r$ respectively. Let the survival function at T_o be $S(T_o)$ which is known. Assume also that $0 \leq t_1 < t_2 < \cdots < t_r$. The likelihood function is then given by

$$L = \prod_{i=1}^{r} p_i \left[S(T_o)\right]^{n-r} . \tag{3.1}$$

Subject to a continuity constraint;

$$\sum_{i=1}^{r} p_i + S(T_o) = 1, \tag{3.2}$$

the likelihood function, L is maximized using the Lagrange multiplier argument as follows: Let

$$Q_1 = \sum_{i=1}^{r} \log p_i + (n-r)\log S(T_o) + \lambda[1 - \sum_{i=1}^{r} p_i - S(T_o)],$$

where λ is the Lagrange multiplier. Differentiating Q_1 with respect to p_i and λ and at the maximum point $dQ_1/dp_i = 0$ for all $i = 1, \ldots, r$ and $dQ_1/d\lambda = 0$, from which we obtain the estimates of $p_i, i = 1, \ldots, r$ and λ. The estimates are $\hat{\lambda} = r/(1 - S(T_o))$ and $\hat{p}_i = (1 - S(T_o))/r), i = 1, \ldots, r$. With the estimates of p_i's and λ we obtain the maximized likelihood as

$$L_{max} = \left(\frac{1 - S(T_o)}{r}\right)^r (S(T_o))^{n-r}.$$

Under the hypothesis, $H_o : T = Median$, we can obtain a median constraint. To do so, we define

$$z_i = \begin{cases} 1 & \text{if } t_i \leq Median \\ -1 & \text{if } t_i > Median \end{cases}$$

and the median constraint is given by

$$\sum_{i=1}^{r} p_i z_i = 0. \tag{3.3}$$

Now, the likelihood ratio for testing H_o is given by

$$R = \frac{L(Median)}{L_{max}} = \left[\frac{r}{1 - S(T_o)}\right]^r \prod_{i=1}^{r} p_i, \tag{3.4}$$

and maximized subject to the constraints (3.2) and (3.3). To employ Lagrange multiplier argument, we let

$$\begin{aligned} Q_2 &= r\log\left(\frac{r}{1 - S(T_o)}\right) + \sum_{i=1}^{r} \log p_i + \gamma\left(1 - \sum_{i=1}^{r} p_i - S(T_o)\right) \\ &- \lambda\left(\sum_{i=1}^{r} p_i z_i\right). \end{aligned} \tag{3.5}$$

Differentiating Q_2 with respect to p_i, γ and λ and set each equation to zero at the maximum point, we obtain

$$p_i = \frac{1}{(\gamma + \lambda z_i)}, \tag{3.6}$$

$$\sum_{i=1}^{r} p_i + S(T_o) = 1, \tag{3.7}$$

and

$$\sum_{i=1}^{r} p_i z_i = 0. \tag{3.8}$$

Solving these equations we obtain the estimate for γ to be

$$\hat{\gamma} = \frac{r}{1 - S(T_o)}. \tag{3.9}$$

Substituting (3.6) and (3.9) into (3.8) and simplifying to obtain

$$\sum_{i=1}^{r} \frac{z_i(1 - S(T_o))}{r + \lambda[z_i(1 - S(T_o))]} = 0. \tag{3.10}$$

The root of (3.10), λ_o may be obtained empirically. The maximized likelihood ratio on substituting the value of λ_o is given by

$$R_o = \prod_{i=1}^{r} \left\{ 1 + \frac{\lambda_o}{r} [z_i(1 - S(T_o))] \right\}^{-1}. \tag{3.11}$$

To establish the following theorem we need the following definitions:

A sequence of random variables $\{T_n\}$, with respective distribution functions $\{F_n\}$, is said to be bounded in probability (which is denoted by $T_n = O_p(1)$) if for every $\epsilon > 0$ there exist M_ϵ and N_ϵ such that

$$F_n(M_\epsilon) - F_n(-M_\epsilon) > 1 - \epsilon \text{ for all } n > N_\epsilon.$$

Let $T_1, T_2, \ldots$ and T be random variables on an appropriate probability space. We say that T_n converges in probability to T (which is denoted by $T_n - T = o_p(1)$) if

$$\lim_{n \to \infty} P(|T_n - T| < \epsilon) = 1, \text{ for every } \epsilon > 0.$$

For detail discussions of the above, the reader may see Serfling (1980).

Theorem 3.1.1 *Define the empirical log-likelihood statistic as*

$$\log R_o = -\sum_{i=1}^{r} \log\left\{1 + \frac{\lambda_o}{r}\left[z_i\left(1 - S(T_o)\right)\right]\right\},$$

where the survival function at T_o, $S(T_o)$ is known. Under proper regularity conditions, minus twice the log-likelihood ratio statistic converges to $\chi^2_{(1)}$ in distribution.

Proof

Recall, we defined

$$z_i = \begin{cases} 1 & \text{if } t_i \leq Median \\ -1 & \text{if } t_i > Median \end{cases}$$

From the above definition, it is easy to show that,

$$E(Z_i) = P(t_i \leq Median) - P(t_i > Median) = 0$$

and

$$Var(Z_i) = E(Z_i)^2 = P(t_i \leq Median) + P(t_i > Median) = 1.$$

To obtain an estimate for λ we solve the following;

$$g(\lambda) = \sum_{i=1}^{r} \frac{z_i(1 - S(T_o))}{r + \lambda[z_i(1 - S(T_o))]} = 0.$$

We may rewrite the above as

$$g(\lambda) = \frac{1}{r}\sum_{i=1}^{r} z_i[1 - S(T_o)] - \frac{\lambda}{r^2}\sum_{i=1}^{r}\{z_i[1 - S(T_o)]\}^2 + O_p(1/r) = 0. \tag{3.12}$$

From this we obtain

$$\lambda_o \doteq \frac{r\sum_{i=1}^{r} z_i}{[1 - S(T_o)]\sum_{i=1}^{r} {z_i}^2}. \tag{3.13}$$

Now, let

$$\begin{aligned} W &= -2\log R_o \\ &= 2\sum_{i=1}^{r} \log\left\{1 + \frac{\lambda_o}{r}[z_i(1 - S(T_o)]\right\}. \end{aligned} \tag{3.14}$$

Expanding (3.14) and simplifying we obtain

$$W = 2\frac{\lambda_o}{r}\sum_{i=1}^{r} z_i[1 - S(T_o)] - \frac{\lambda_o^2}{r^2}\sum_{i=1}^{r}\{z_i[1 - S(T_o)]\}^2 + o_p(1), \tag{3.15}$$

for sufficiently large r. We note that $|\frac{\lambda_o}{r}[z_i(1-S(T_o))]| < 1$. Substituting (3.13) into (3.15) to obtain

$$\begin{aligned} W &= 2\frac{(\sum_{i=1}^r z_i)^2}{\sum_{i=1}^r {z_i}^2} - \frac{(\sum_{i=1}^r z_i)^2}{\sum_{i=1}^r {z_i}^2} + o_p(1) \\ &= \frac{(\sum_{i=1}^r z_i)^2}{\sum_{i=1}^r {z_i}^2} + o_p(1) \\ &= \frac{r\bar{z}^2}{S^2} + o_p(1) \qquad \rightarrow \chi^2_{(1)}. \end{aligned}$$

Since $\frac{\sqrt{r}\bar{z}}{S}$ is asymptotically standard normally distributed, by the central limit theorem where

$$\bar{z} = \frac{1}{r}\sum_{i=1}^r z_i \text{ and } S^2 = \frac{1}{r}\sum_{i=1}^r {z_i}^2.$$

Hence $W \rightarrow \chi^2_{(1)}$.

For the general case of the above result, where $S(T_o)$ is unknown, we turn to the next section.

3.2 Empirical Likelihood Inference on the Median when $S(T_o)$ is unknown

Let $t_1, \ldots, t_r$ denote once again the distinct times of observed failures (which may be discrete or may be made so by grouping) before the fixed time T_o with probabilities $p_1, \ldots, p_r$ respectively. Let the survival function at T_o be $S(T_o)$ which is unknown. Let $\hat{S}(T_o)$ be the Kaplan-Meier estimate at $T = T_o$. Again we assume that $0 \leq t_1 < t_2 < \cdots < t_r$. The likelihood function is therefore given by

$$L = \prod_{i=1}^r p_i [S(T_o)]^{n-r}.$$

Following the arguments given in the previous section, we can show that the likelihood ratio, when $S(T_o)$ is replaced by the Kaplan-Meier estimate at $T = T_o$, $\hat{S}(T_o)$, is given by

$$\hat{R} = \left[\frac{r}{1-\hat{S}(T_o)}\right]^r \prod_{i=1}^r \hat{p}_i, \tag{3.16}$$

where

$$\hat{p}_i = \frac{1-\hat{S}(T_o)}{r\left[1+\frac{\hat{\lambda}_o}{r}z_i\left(1-\hat{S}(T_o)\right)\right]},$$

and

$$\hat{\lambda}_o = \frac{r}{1 - \hat{S}(T_o)} \frac{\sum_{i=1}^{r} z_i}{\sum_{i=1}^{r} z_i^2}.$$

Substituting these estimates into (3.16) and simplifying we obtain

$$\hat{R} = \prod_{i=1}^{r} \left\{ 1 + z_i \frac{\sum_{i=1}^{r} z_i}{\sum_{i=1}^{r} z_i^2} \right\},$$

which is independent of $\hat{S}(T_o)$. Now, let

$$\begin{aligned} \hat{W} &= -2 \log \hat{R} \\ &= 2 \sum_{i=1}^{r} \log \left(1 + z_i \frac{\sum_{i=1}^{r} z_i}{\sum_{i=1}^{r} z_i^2} \right). \end{aligned} \tag{3.17}$$

Expanding (3.17) and simplifying, we obtain

$$\hat{W} = \frac{\left(\sum_{i=1}^{r} z_i\right)^2}{\sum_{i=1}^{r} z_i^2} + o_p(1),$$

which is distributed as $\chi^2_{(1)}$. The next section discusses the various methods of interval estimation by Brookmeyer and Crowley (1982), Emerson (1982) and that of the empirical likelihood ratio mentioned in Sections (3.1) and (3.12).

3.3 Interval Estimation for the Median

We will consider first the interval estimation using Brookmeyer and Crowley (1982) method. In their method, an asymptotic $1 - \alpha$ confidence region R_α was considered. This is obtained as the set of all parameter values not rejected by the sign test at level α. That is,

$$R_\alpha = \left\{ m | (\hat{S}^o(m) - \frac{1}{2})^2 \leq C_\alpha \left(\hat{S}^o(m) \right)^2 \sum_{\substack{\text{distinct} \\ X_i \leq M}} \frac{d_i}{N_x(X_i)\{N_x(X_i) + d_i\}} \right\}.$$

To find the region R_α, it is necessary only to check if observed death times are in the region. This is because the Kaplan-Meier estimate and its estimated variance jump only at observed death times. Thus, if t_1 and t_2 are two consecutive observed death times with $t_1 < t_2$ and $t_1 \epsilon R_\alpha$, then the interval $[t_1, t_2)$ is contained in R_α. One would expect the confidence region R_α to be an interval which includes the estimated

median. This is generally true. Simulation results suggest that the confidence region is almost always an interval. Thus, it is reasonable to consider only the interval part of the confidence region R_α. Now let $\{t_i\}$ be the observed distinct death times. The confidence interval is defined as $I_\alpha = [t_i, t_j)$, where t_i is the smallest observed death time in R_α with $\hat{S}^o(t_i) > .5$, and t_j is the smallest observed death time not in R_α with $\hat{S}^o(t_j) < .5$.

It must be mentioned that occassionally it happens that an upper confidence limit cannot be obtained. If the last observed death time is in I_α, then I_α becomes a one-sided confidence interval of the form $[t_i, \infty)$. Furthermore, if the Kaplan-Meier survival curve does not reach the median because of extensive censoring, only a lower confidence limit (or perhaps an empty interval) can be obtained.

The second method we would discuss is by Emerson (1982), which is simply a generalization of the sign test method. A nonparametric confidence interval is construted using the following steps:

1. Construction of $\hat{S}$, the product-limit estimate of the survival function.

2. Tests of hypotheses of the form $H_o : median = t_j$, where t_j is an observed failure time. The level-α test is based upon the statistic $\hat{N}_*$, which is computed from $\hat{S}$; its distribution is approximated by the binomial distribution with parameters n and $\frac{1}{2}$, using interpolation.

3. Inversion of the tests to provide a two-sided confidence interval for the median with confidence coefficient $1 - \alpha$.

Finally, Interval estimates for the median using either W or $\hat{W}$ depending on the underlying assumptions. It must be emphasized that obtaining interval estimates for the median using either of the above results is analytically not feasible. We therefore employ a computational approach. The approach is similar to the one discussed under Brookmeyer and Crowley (1982). In each of the cases, since $\mathrm{P}(\chi^2_{(1)} \leq \chi^2_{(1),\alpha}) = (1-\alpha)$, a $(1-\alpha)*100\%$ confidence interval for the median is found as the set of all parameter (median) values giving $W \leq \chi^2_{(1),\alpha}$, where $\chi^2_{(1),\alpha}$ is the αth percentile of the $\chi^2_{(1)}$ distribution. By computing W for several values of the parameter (median), we can locate the "exact" values of the median that make $W = \chi^2_{(1),\alpha}$. The two values that give the most approximate $\chi^2_{(1),\alpha}$ are the approximate $1 - \alpha$ confidence interval for

the median. It must be emphasized that this method can be used to obtain one-tail confidence intervals when it is required, whereas in the previous approaches, one-tail confidence intervals are determined by the nature of censoring or the nature of the data under consideration. The next section considers some of the simulation results of this study.

3.4 Simulation Study

In this section we would describe simulation experiments carried out to compare empirical likelihood ratio confidence intervals and the observed coverage probabilities in the hypothesis testing of the median using W with that of other nonparametric methods proposed by Brookmeyer and Crowley (1982) and Emerson (1982).

For the data to be used in the simulation study, we consider the one-parameter exponential distribution and the Weibull distribution. We generated 1000 samples of 25 exponential random variables and 1000 samples of 50 exponential random variables each with parameter .5. Also 1000 samples of 25 Weibull random variables and 1000 samples of 50 Weibull random variables each with scale parameter 1 and shape parameter .5. Additional Weibull random variables are generated with the same scale parameter and shape parameters, .2, .8, 1.0, 1.2 and 1.5 respectively. Each of these random variables is considered as a lifetime observation. The random samples were generated using the software Splus. However, all the simulation programs were run in the C programming language.

We now describe how the data were generated appropriately for the simulation study. The observations from the exponential distribution are used here as an example. Suppose there are 20 percent censored observations, then since it is assumed that all the censored observations occur after the censored T_o, it is easy to find that the appropriate quantile that would give rise to the 20 percent censoring is $T_o = 3.218876$ for each of the samples. Thus we have an expected 80 percent uncensored observations before T_o which we consider as counting data and an expected 20 percent censored observations after T_o, which we consider as the known $S(T_o)$. The above is repeated for 10, 30, 35 and 40 percent censoring of the data. We repeat the whole process using the Weibull distribution.

For the nonparametric procedures, the uncensored observations represent the

available data for the simulation. This is so because in both methods by Emerson (1982) and Brookmeyer and Crowley (1982), we know that estimation or inference are done using only the actually observed failure time observations which in our case are simply the uncensored observations described above.

For every sample generated either for the purpose of the empirical likelihood ratio method or for the nonparametric methods, observed coverage probabilities and the average confidence length at 95% confidence level were recorded. Table 3.1 shows the observed coverage probabilities and average confidence length for the three methods using the exponential distribution, Table 3.2 shows the coverage probabilities of the three methods using the Weibull distribution under the various shape parameters mentioned and Table 3.3 shows the corresponding average confidence lengths.

Table 3.1: Oberved Coverage Probabilities and average interval length or Exponential (.5) survival distribution: 1-α =.95

Sample size		25		50	
Method	Percent censored	Coverage	Length	Coverage	Length
Empirical Likelihood					
Case I	10	.949	1.52	.964	1.11
	20	.947	1.38	.958	1.05
	30	.943	1.17	.950	.86
	35	.938	1.01	.958	.75
	40	.919	.87	.934	.68
Emerson	10	.973	2.09	.953	1.31
	20	.973	2.09	.953	1.31
	30	.973	1.80	.953	1.31
	35	.973	1.55	.953	1.19
	40	.973	1.33	.953	.97
Brookmeyer & Crowley	10	.956	1.95	.941	1.15
	20	.956	1.95	.941	1.15
	30	.956	1.95	.941	1.15
	35	.956	1.95	.941	1.15
	40	.956	1.95	.941	1.15

The results under Brookmeyer and Crowley's method are the same at all the specified levels of censoring because their statistic is independent of right censoring beyond the median.

Table 3.2: Observed coverage probabilities for Weibull survival distribution: $1\text{-}\alpha = .95$

Sample size		25					50				
	Percent	*Shape parameter*, β					*Shape parameter*, β				
Method	censored	.5	.8	1.0	1.2	1.5	.5	.8	1.0	1.2	1.5
Empirical likelihood											
Case I	10	.940	.941	.954	.954	.947	.963	.946	.949	.950	.964
	20	.946	.945	.955	.958	.949	.962	.949	.954	.951	.958
	30	.941	.950	.952	.965	.944	.952	.946	.954	.951	.942
	35	.938	.940	.948	.951	.933	.956	.947	.947	.947	.942
	40	.935	.925	.934	.928	.922	.953	.944	.938	.942	.941
Emerson	10	.972	.973	.975	.969	.974	.960	.951	.950	.942	.967
	20	.972	.973	.975	.969	.974	.960	.951	.950	.942	.967
	30	.972	.973	.975	.969	.974	.960	.951	.950	.942	.967
	35	.972	.973	.975	.969	.974	.960	.951	.950	.942	.967
	40	.972	.973	.975	.969	.974	.960	.951	.950	.942	.967
Brookmeyer &	10	.956	.955	.971	.953	.959	.951	.941	.952	.941	.952
Crowley	20	.956	.955	.971	.953	.959	.951	.941	.952	.941	.952
	30	.956	.955	.971	.953	.959	.951	.941	.952	.941	.952
	35	.956	.955	.971	.953	.959	.951	.941	.952	.941	.952
	40	.956	.955	.971	.953	.959	.951	.941	.952	.941	.952

The results in Table 3.2 using Emerson's method and that of Brookmeyer and Crowley can be seen to be the same at all the specified levels of censoring. This is because their methods are independent of right censoring beyond the median.

Table 3.3: Average Confidence interval length for Weibull survival distribution

Sample size		25					50				
	Percent	*Shape parameter*, β					*Shape parameter*, β				
Method	censored	.5	.8	1.0	1.2	1.5	.5	.8	1.0	1.2	1.5
Empirical likelihood											
Case I	10	1.23	.89	.75	.66	.53	.80	.64	.54	.48	.39
	20	1.03	.78	.67	.59	.50	.69	.56	.48	.43	.36
	30	.81	.64	.57	.49	.43	.58	.57	.40	.34	.31
	35	.58	.56	.48	.44	.39	.50	.42	.36	.32	.27
	40	.55	.47	.42	.37	.28	.40	.34	.30	.26	.22
Emerson	10	1.87	1.22	1.04	.90	.76	1.03	.76	.65	.56	.47
	20	1.87	1.22	1.04	.90	.76	1.03	.76	.65	.56	.47
	30	1.47	1.04	.90	.79	.67	1.03	.76	.65	.56	.47
	35	1.18	.87	.78	.69	.59	.90	.68	.59	.51	.43
	40	.93	.73	.67	.59	.52	.68	.55	.48	.42	.36
Brookmeyer &	10	1.50	1.04	.86	.76	.63	.81	.64	.56	.49	.40
Crowley	20	1.50	1.04	.86	.76	.63	.81	.64	.56	.49	.40
	30	1.50	1.04	.86	.76	.63	.81	.64	.56	.49	.40
	35	1.50	1.04	.86	.76	.63	.81	.64	.56	.49	.40
	40	1.50	1.04	.86	.76	.63	.81	.64	.56	.49	.40

We observe in Table 3.1 that the three methods show consistently excellent coverage probabilities and there is no dramatic change of these values as the percentage of censoring increases. Among the three methods we also observe that the average interval lengths of the confidence intervals are somewhat comparable for the empirical method and the method by Brookmeyer and Crowley which are shorter than the method by Emerson. We also observe that the average lengths tend to decrease at higher percentages of censoring for the method by Emerson. A similar occurrence is exhibited by the Case I method of the empirical likelihood method. Table 3.2 shows the coverage probabilities again of the three methods but here under the Weibull distribution with different shape parameters. Again there seem to be a consistently very good coverage probabilities for all the methods. The methods seem to differ in

terms of their average interval lengths, with the empirical likelihood and Brookmeyer and Crowley methods having an edge. The results of the average interval lengths are shown in Table 3.3.

One is therefore left with a choice of which method to use for survival data analysis which may require testing or constructing confidence intervals for the median survival time or generally for any quantile survival time. When one's motivation is to obtain shorter interval lengths with good coverage probabilities then we may recommend the use of either the empirical likelihood method or the method by Brookmeyer and Crowley but this would be at a higher cost of computation. Otherwise any one of the methods seem quite appropriate in such analysis.

It is interesting also to note that we deal only with the first order properties of the empirical likelihood estimators; higher order properties need study.

Chapter 4

Estimation in Survey Sampling in the Presence of Auxiliary Information

In sampling from a finite population it is usual practice to find some supplementary information for various units comprising the population. Such information is generally based on previous censuses or large-scale surveys. For example, the number of inhabitants in different villages may be known from previous census of population. This available information is called auxiliary information and it may be used in many ways such as at the estimation stage, or at the selection stage or at both stages. In a sample selection, it may be used by selecting the sample with probability proportional to the value of the auxiliary information. Ratio method of estimation, difference method of estimation, regression method of estimation and empirical likelihood method of estimation are some examples of the use of auxiliary information at the estimation stage. It is important to note that the auxiliary information is used primarily for improving the precision of estimates. Our discussion will focus primarily on the use of auxiliary information in estimation under empirical likelihood framework. We would briefly review the conventional methods mentioned above in the subsequent sections of this chapter.

4.1 Ratio Method of Estimation

In order to obtain ratio estimate of the population total of the study variable, Y of y_i, an auxiliary variate x_i, correlated with y_i is obtained for each unit in the sample. The population total X of the x_i must also be known. In practice, x_i is often the value of y_i at some previous time when a complete census was taken. The aim of this method is to obtain increased precision by taking advantage of the correlation between y_i and x_i.

Suppose the ratio $R = Y/X$, where Y and X are the population totals for the variables y and x, is to be estimated on the basis of a sample selected through any given sampling scheme. Let $\hat{Y}$ and $\hat{X}$ be unbiased estimators of Y and X respectively.

Then an estimator of the ratio R is given by

$$\hat{R} = \frac{\hat{Y}}{\hat{X}}. \tag{4.1}$$

Similarly, a ratio estimator of Y is given by

$$\hat{Y}_R = \hat{R}X = \frac{\hat{Y}}{\hat{X}}X, \tag{4.2}$$

when information on the total of X of a related auxiliary variable x is available. It is noted that for estimating a total by the method of ratio estimation we should know the value of X from other source.

For the commonly used selection procedures, the ratio estimator given above is, in general, biased for the corresponding population ratio. However, it is to be noted that a biased estimator may be prefferred to an unbiased estimator, if the mean square error of the former is less than the variance of the latter. Under simple random sampling, the expressions for bias and mean square error (MSE) for sufficiently large sample size are given below. An approximate expression for the bias is given by

$$B(\hat{R}) = E(\hat{R} - R) \doteq \frac{1}{X^2}\left[RV(\hat{X}) - Cov(\hat{X}, \hat{Y})\right]. \tag{4.3}$$

The bias of the ratio estimator can be estimated by substituting in the above equation estimators of X, R, $V(\hat{X})$ and $Cov(\hat{X}, \hat{Y})$. Thus, an estimator of $B(\hat{R})$ is given by

$$b(\hat{R}) = \frac{1}{\hat{X}^2}\left[\hat{R}v(\hat{X}) - cov(\hat{X}, \hat{Y})\right].$$

We now turn our attention to the MSE of the ratio estimator. Since the ratio estimator is biased, we have to consider its mean square error for the purpose of comparing its efficiency with that of any other estimator. The MSE is given by

$$\begin{aligned} M(\hat{R}) &= E(\hat{R} - R)^2 \\ &\doteq \frac{1}{X^2}\left[V(\hat{Y}) - 2RCov(\hat{X}, \hat{Y}) + R^2V(\hat{X})\right]. \end{aligned} \tag{4.4}$$

For detail derivation of the above results, the reader may refer to Cochran (1977). As in the case of bias, it is possible to estimate the MSE by simply substituting estimators of X, R, $V(\hat{X})$, $V(\hat{Y})$ and $Cov(\hat{X}, \hat{Y})$ in the above equation. Thus we have

$$v(\hat{R}) = \frac{1}{\hat{X}^2}\left[v(\hat{Y}) - 2\hat{R}cov(\hat{X}, \hat{Y}) + \hat{R}^2v(\hat{X})\right].$$

The bias of $\hat{Y}_R$ is obtained simply by multiplying the bias of $\hat{R}$ by X and the MSE of $\hat{Y}_R$ is $X^2V(\hat{R})$. The question is asked whether the ratio estimate will always be an improvement on the simple average of the survey variates. The answer is that the improvement depends on the strength of correlation (positive) between the survey variates and the auxiliary variates. The higher the strength of correlation between the survey variates and the auxiliary variates, the better the ratio method of estimation would be over the simple average of the survey variates, otherwise, there is not much to be gained.

4.2 Regression Method of Estimation

Under the ratio method of estimation, we considered the question of improving the conventional unbiased estimator $\hat{Y}$ by multiplying it with the factor $X/\hat{X}$, where $\hat{X}$ is an unbiased estimator of the total of a suitably chosen supplementary variable. Here we examine the possibility of improving upon $\hat{Y}$ by considering the estimator of the form

$$\hat{Y}(k) = \hat{Y} + k(X - \hat{X}), \tag{4.5}$$

where k is selected optimally by minimizing the variance of $\hat{Y}(k)$. The resulting estimator is the *regression estimator* given by

$$\hat{Y}'_r = \hat{Y} + \beta(X - \hat{X}), \tag{4.6}$$

and the procedure of estimation is known as *regression method of estimation.* Its variance is given by

$$V(\hat{Y}'_r) = V(\hat{Y})\left\{1 - \rho^2(x, y)\right\}, \tag{4.7}$$

where $\rho(x, y)$ is the correlation coefficient between x and y and β is the regression coefficient of y on x. In actual practice, the exact value of β may not be known and it may have to be estimated on the basis of a sample. If $\hat{\beta}$ is an estimator of β, we get

$$\hat{Y}_r = \hat{Y} + \hat{\beta}(X - \hat{X}). \tag{4.8}$$

This estimator is generally biased for Y and its bias is given by

$$B(\hat{Y}_r) = E(\hat{Y}_r) - Y \doteq -Cov(\hat{X}, \hat{\beta}). \tag{4.9}$$

The variance of $\hat{Y}_r$ to a first order of approximation is given by

$$V(\hat{Y}_r) = V(\hat{Y}) - 2\beta Cov(\hat{X}, \hat{Y}) + \beta^2 V(\hat{X}), \tag{4.10}$$

which reduces to (4.6).

A consistent estimator of the variance may be obtained by substituting estimators of $V(\hat{Y})$ and $\rho(x, y)$ in the variance expression. The regression estimator is not commonly used in practice due to the fact that the calculation of the estimate of the regression coefficient in large-scale surveys becomes cumbersome and time consuming. Further, since the regression line passes through the origin or close to the origin in most cases usually met with in practice, the ratio estimator is generally used instead of the more complicated regression estimator. Again detail discussions of the above may be found in any of the standard texts in sampling.

4.3 Difference Method of Estimation

Suppose from a simple random sample, we obtain unbiased estimators $\hat{Y} = N\bar{y}$ and $\hat{X} = N\bar{x}$ of Y and X, respectively, then an unbiased estimate of Y, given by

$$\hat{Y}_D = \hat{Y} + (X - \hat{X}), \tag{4.11}$$

is known as a difference estimate whose variance is

$$V(\hat{Y}_D) = V(\hat{Y}) - 2Cov(\hat{X}, \hat{Y}) + V(\hat{X}). \tag{4.12}$$

An important advantage of $\hat{Y}_D$ is that it is particularly simple and always unbiased.

4.4 Empirical Likelihood Method of Estimation

Suppose $p_i(i\epsilon s)$ is the probability mass of sampling the survey variate $Y_i(i\epsilon s)$ from the population $\prod$. Where s is the sample of interest. Suppose also that there is auxiliary information x_i for each of the survey variate sampled. We can obtain estimates of population characteristics using empirical likelihood method as follows. Suppose we

are particularly interested in obtaining estimate of the mean of Y, then we aim at maximizing

$$\prod_{i \in s} p_i,$$

subject to the constraints

$$\sum_{i \in s} p_i = 1, \quad \sum_{i \in s} w(x_i) p_i = 0 \quad (0 \leq p_i \leq 1). \tag{4.13}$$

When the available information is $\bar{X}$, $w(x)$ takes the form $w(x) = x - \bar{X}$; when the information is given as the population median m_X, $w(x)$ takes the form $w(x) = I_{[x \leq m_X]} - .5$. Chen and Qin (1993) show that almost any information about X may be used to improve the estimation of the population characteristics of Y.

Now from (4.13) and employing the Lagrange multiplier argument we obtain

$$p_i = \frac{1}{n[1 + \lambda w(x_i)]},$$

$$\sum_{i \in s} \frac{w(x_i)}{1 + \lambda w(x_i)} = 0, \tag{4.14}$$

from which we may solve for λ and therefore obtaining estimates of p_i. The estimate of the mean of Y is then given by

$$\bar{y}^{(E)} = \sum_{i \in s} y_i \hat{p}_i. \tag{4.15}$$

This estimator is biased and for sufficiently large n an approximate expression for the variance is given by

$$V(\bar{y}^{(E)}) = \frac{1}{n} \left[S_y^2 - \frac{S_{yw}^2}{S_w^2} \right], \tag{4.16}$$

where

$$S_y^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \bar{Y})^2, \quad S_w^2 = \frac{1}{N-1} \sum_{i=1}^{N} w^2(x_i) \text{ and}$$

$$S_{yw} = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \bar{Y}) w(x_i).$$

For detailed derivation of the asymptotic variance of $\bar{y}^{(E)}$, the reader may refer to Chen and Qin (1993). We observe that in all the above methods of estimation we

have assumed that auxiliary information is available in addition to known value of $\bar{X}$. However, there are situations with unknown $\bar{X}$, but information on the variable x may be less expensive to obtain compared to that of the study variable y. In such cases, the usual thing to do is to make recourse to the technique of double sampling, leading to a natural modification of the above methods of estimation. This technique ie reviewed in the next section.

Olkin (1958) extended the ratio estimate to the situation in which p auxiliary x-variables $(x_1, x_2, \ldots, x_p)$ are available. For further discussions on this, the reader may see Olkin (1958).

4.5 Double Sampling

As we have seen from previous disussions, a number of methods of estimation in sampling like the ratio method of estimation, the regression method of estimation and the empirical likelihood method of estimation depend on the possession of advance information about an auxiliary variate x and its population mean $\bar{X}$. When such information is lacking, it is sometimes convenient and relatively cheap to take a large preliminary sample in which x alone is measured. The purpose of this sample is to furnish a good estimate of $\bar{X}$. In a survey whose function is to make estimates for some other variate y, it may pay to devote part of the resources to this preliminary sample, although this means that the size of the sample in the main survey on y must be decreased. This technique is known as *double sampling* or *two-phase sampling.*

In double sampling, first-phase sampling information may be used as supplementary information in order to improve the accuracy of second-phase information, by the same methods, ratio, regression and empirical, that are applicable where supplementary information on the whole population is available. Thus, in a crop estimation survey based on farms as sampling units, relatively large sample of farms may be taken for the determination of the acreage of the crop, and the yields may be determined on a sub-sample only of these farms.

We must mention that two-phase sampling technique is profitable only if the gain in precision from ratio, regression, difference or empirical likelihood estimates more than offsets the loss in precision due to reduction in the main sample. How this technique is employed in the presence of a second auxiliary information is discussed

in the next section.

4.6 Estimation in Two-Phase Sampling using Two Auxiliary Variables

For this section and subsequent ones, we would employ a finite population $U = \{1, 2, \ldots, k, \ldots, N\}$. Suppose y and x are the variable for study and the auxiliary variable, taking values y_k and x_k respectively for the kth unit such that the two variables are strongly related but no information is available on the population mean $\bar{X}$ of x, we aim at estimating the population mean $\bar{Y}$ of y from a second sample s_2, obtained through a two-phase selection. We use simple random sampling without replacement in selecting sample $s_1 (s_1 \subset U)$ and sample s_2 from s_1 with sizes n and m respectively. We note that for the units in sample s_1 we observe only x which enables us to obtain a good estimate of $\bar{X}$ and for the units in s_2 we observe y also.

Under these sampling schemes, we let

$$\bar{x}_{s_2} = \frac{1}{m}\sum_{k \epsilon s_2} x_k, \quad \bar{y}_{s_2} = \frac{1}{m}\sum_{k \epsilon s_2} y_k \quad \text{and} \quad \bar{x}_{s_1} = \frac{1}{n}\sum_{k \epsilon s_1} x_k,$$

such that the sampling ratio and regression estimators are given respectively as

$$t_R = \bar{y}_{s_2}\frac{\bar{x}_{s_1}}{\bar{x}_{s_2}} \tag{4.17}$$

and

$$t_{RG} = \bar{y}_{s_2} + b_{xy}^{(s_2)}(\bar{x}_{s_1} - \bar{x}_{s_2}), \tag{4.18}$$

where $b_{xy}^{(s_2)}$ is the sample regression coefficient of y on x computed using data on s_2. To a first order of approximation, the estimators above have mean square errors given by

$$M(t_R) = f_{s_2}\left(S_y^2 - 2R_{xy}\rho_{xy}S_yS_x + R_{xy}^2S_x^2\right) + f_{s_1}\left(2R_{xy}\rho_{xy}S_yS_x - R_{xy}^2S_x^2\right)$$

and

$$M(t_{RG}) = S_y^2\left[f_{s_2}(1-\rho_{xy}^2) + f_{s_1}\rho_{xy}^2\right],$$

where

$$f_{s_2} = \left(\frac{1}{m} - \frac{1}{N}\right), \quad f_{s_1} = \left(\frac{1}{n} - \frac{1}{N}\right), \quad S_y^2 = \frac{1}{N-1}\sum_{k=1}^{N}(y_k - \bar{Y})^2,$$

$$S_x^2 = \frac{1}{N-1}\sum_{k=1}^{N}(x_k - \bar{X})^2, \quad R_{xy} = \frac{\bar{Y}}{\bar{X}},$$

and ρ_{xy} is the correlation coefficient between y and x.

Use of a Second Auxiliary Variable

Sometimes even if $\bar{X}$ is unknown, information on a cheaply ascertainable variable z, closely related to x but compared to x remotely related to y, is available on all units of the population (e.g. y is value of cattle and/or calves sold live in 1964, x is the number of cattle and/or calves sold in 1964 and z is the number of farms reporting sale of cattle and/or calves sold live in 1964, $\rho_{xy} \geq \rho_{yz}$). Chand (1975), Kiregyera (1980, 1984), Prasad and Srivenkataramana (1980), Srivastava et al. (1988, 1990) gave brief discussions of this type of situation.

Chand (1975) suggested a chain ratio-type estimator given by

$$t_{11} = \bar{y}_{s_2} \frac{\bar{x}_{s_1} \bar{Z}}{\bar{x}_{s_2} \bar{z}_{s_1}}, \tag{4.19}$$

where $\bar{Z}$ is the known population mean of z and $\bar{z}_{s_1} = \sum_{k \epsilon s_1} z_k/n$. The MSE of this estimator to a first order of approximation is given by

$$\begin{aligned} M(t_{11}) = f_{s_2} \left[S_y^2 - 2R_{xy}\rho_{xy}S_xS_y + R_{xy}^2 S_x^2\right] &+ f_{s_1} \left[R_{yz}^2 S_z^2 - 2R_{yz}\rho_{yz}S_yS_z \right. \\ &+ \left. 2R_{xy}\rho_{xy}S_xS_y - R_{xy}^2 S_x^2\right]. \end{aligned} \tag{4.20}$$

It is easy to see that Chand obtained t_{11} from simply replacing $\bar{x}_{s_1}$ by the ratio estimator

$$t_R^{(s_1)} = \frac{\bar{x}_{s_1} \bar{Z}}{\bar{z}_{s_1}}.$$

After the work by Chand (1975), Kiregyera (1980) proposed an estimator which he refers to as ratio-to-regression estimator

$$t_{12} = \frac{\bar{y}_{s_2}}{\bar{x}_{s_2}} \left[\bar{x}_{s_1} + b_{xz}^{(s_1)}(\bar{Z} - \bar{z}_{s_1})\right], \tag{4.21}$$

obtained simply by replacing $\bar{x}_{s_1}$ by the regression estimator

$$t_{RG}^{(s_1)} = \bar{x}_{s_1} + b_{xz}^{(s_1)}(\bar{Z} - \bar{z}_{s_1}),$$

where $b_{xz}^{(s_1)}$ is the sample regression coefficient of x and z based on s_1. The MSE of t_{12} to $O(\frac{1}{m})$ is

$$\begin{aligned} M(t_{12}) &= f_{s_2} \left[S_y^2 - 2R_{xy}\rho_{xy}S_xS_y + R_{xy}^2 S_x^2\right] + f_{s_1} \left[2R_{xy}\rho_{xy}S_yS_x - R_{xy}^2 S_x^2 \right. \\ &+ \left. R_{xy}R_{yz}\rho_{xz}S_xS_z - R_{xy}\rho_{yz}\rho_{xz}S_xS_y\right]. \end{aligned} \tag{4.22}$$

Kiregyera (1984) also extended this formulation to develop a ratio-in-regression estimator given by

$$t_{21} = \bar{y}_{s_2} + b_{xy}^{(s_2)} \left(\frac{\bar{x}_{s_1}}{\bar{z}_{s_1}} \bar{Z} - \bar{x}_{s_2} \right), \tag{4.23}$$

and a regression-in-regression estimator given by

$$t_{22} = \bar{y}_{s_2} + b_{xy}^{(s_2)} \left[(\bar{x}_{s_1} + b_{xz}^{(s_1)} (\bar{Z} - \bar{z}_{s_1})) - \bar{x}_{s_2} \right]. \tag{4.24}$$

The MSE's of t_{21} and t_{22} to $O(\frac{1}{m})$ are respectively,

$$M(t_{21}) = \left[\frac{1}{m} \alpha_1 + \frac{1}{n} \alpha_2 \right], \tag{4.25}$$

where

$$\alpha_1 = (1 - \rho_{xy}^2) S_y^2 \text{ and } \alpha_2 = \frac{\bar{X}}{\bar{Y}\bar{Z}^2} \frac{S_{xy}^3 S_z^2}{S_x^4} - \frac{2 S_{xy}^2 S_{yz}}{\bar{Y}\bar{Z} S_x^2},$$

and

$$\begin{aligned} M(t_{22}) &= \frac{1}{m} S_y^2 - \left(\frac{1}{m} - \frac{1}{n} \right) \rho_{xy}^2 S_y^2 + \frac{1}{n R_{xz}} \frac{S_{xz}}{S_z^2 S_{xy}} \left[R_{xy}^2 S_x^2 S_{xy} \right. \\ &+ \left. R_{yz}^2 S_z^2 S_{xy} - 2 R_{yz} R_{xy}^2 S_{xz} S_x^2 \right], \end{aligned} \tag{4.26}$$

where

$$S_{xz} = \frac{1}{N-1} \sum_{k=1}^{N} (x_k - \bar{X})(z_k - \bar{Z}), \quad S_{xy} = \frac{1}{N-1} \sum_{k=1}^{N} (x_k - \bar{X})(y_k - \bar{Y})$$

$$\text{and } S_{yz} = \frac{1}{N-1} \sum_{k=1}^{N} (y_k - \bar{Y})(z_k - \bar{Z}).$$

It can be seen that no general conclusion can be drawn from Kiregyera (1980, 1984) on the comparison of different estimators under mean square error. This is due to the fact that it is almost impossible to check the conditions of preference of t_{12} over t_{11} and t_{22} over t_{21} in practice. We however, note that the conditions of preference of t_{12} over t_R and t_{22} over t_{RG} are respectively

$$\rho_{yz} > \frac{1}{2} \rho_{xz} \frac{C(X)}{C(Y)}$$

and

$$\rho_{yz} > \frac{1}{2} \rho_{xy} \rho_{xz},$$

where $C(X) = S_x/\bar{X}$ and $C(Y) = S_y/\bar{Y}$; ρ_{yz} andρ_{xz} have their usual meanings.

We observe that, when z is remotely related to y, as supposed by Kiregyera, these conditions cannot easily be realized in practice. Therefore, t_{12} or t_{22} may not be efffectively used in many situations. It is also observed that while selecting these alternative estimators for $\bar{x}_{s_1}$, which are $t_R^{(s_1)}$ and $t_{RG}^{(s_1)}$ into the standard ratio and regression estimators, only the relation between x and z was taken into consideration. Therefore, the use of estimators like $t_R^{(s_1)}$ and $t_{RG}^{(s_1)}$ cannot always provide better results.

It is important to note that, for the trivariate distribution under consideration, the correlation between (x, z) is also influenced by the correlation between (x, y) and (y, z). However, these were not taken into consideration by Chand and Kiregyera.

Sahoo et al. (1994) propose an alternative method of estimation of the population mean of the survey variate, y which takes into consideration the influence of the correlation between (x, y) and that between (y, z) on (x, z). Estimates obtained using their method tend to be more efficient than the estimates obtained using the methods by Chand and Kiregyera. The following sub-section discusses the methods by Sahoo et al. (1994).

Sahoo et al. Alternative Approach

The estimators proposed by Sahoo et al. (1994) are

$$t_1 = t_R + b_{yz}^{(s_2)}(\bar{Z} - \bar{z}_{s_1}) \tag{4.27}$$

and

$$t_2 = t_{RG} + b_{yz}^{(s_2)}(\bar{Z} - \bar{z}_{s_1}). \tag{4.28}$$

They referred to these estimators as ratio-cum-regression and regression-cum-regression estimators respectively. The MSE's of these estimators to a first order of approximation are given by

$$M(t_1) = M(t_R) - f_{s_1}\rho_{yz}^2 S_y^2 \tag{4.29}$$

and

$$M(t_2) = M(t_{RG}) - f_{s_1}\rho_{yz}^2 S_y^2. \tag{4.30}$$

For detail derivations of the above results, the reader may see Sahoo et al. (1994). They compared the mean square errors of their estimators with that of the other estimators, namely, the standard ratio estimator, the standard regression estimator, the ratio-type estimator by Chand (1975), the ratio-to-regression estimator by Kiregyera (1980), the ratio-in-regression and the regression-in-regression estimators also by Kiregyera (1984). They came out with the conlusion that among all the competing estimators namely, t_R, t_{11}, t_{12}, t_{RG}, t_{21}, t_{22}, t_1 and t_2, t_2 emerged as the most efficient.

In the next Chapter we propose empirical likelihood estimation in two-phase sampling using two auxiliary variables.

Chapter 5

Empirical Likelihood Estimation in Two-Phase Sampling Using Two Auxiliary Variables

As pointed out by Sahoo et al. (1994), the estimators proposed by Chand (1975) and Kiregyera (1980, 1984) ignored the fact that in the trivariate distribution under consideration, the correlation between (x, z) is influenced by the correlation between (x, y) and that between (y, z) and because of that the estimators proposed by Chand and Kiregyera are not as efficient as those proposed by Sahoo et al. (1994). Sahoo et al. also observed that t_1, is uniformly better than its competitors t_R, t_{11} and t_{12} and similarly, the estimator t_2 is more efficient than its competitors t_{RG}, t_{21} and t_{22}. Also the estimator t_2 emerged as the most efficient among all the estimators discussed in that Chapter.

In this Chapter, we propose two alternative estimators and like Sahoo et al. (1994), the influence of the correlation between one pair of variates on that of the others are taken into consideration in constructing these estimators. Empirical likelihood method of estimation is applied to a modified form of ratio and/or regression estimators. We refer to the resulting estimators as empirical likelihood regression-thru-ratio and empirical likelihood regression-thru-regression estimators.

5.1 Regression-thru-ratio Method of Estimation

We must mention that the population U and the samples s_1 and s_2 with their characteristics described in Chapter 4 are still used here. Let

$$\bar{z}_{s_1} = \frac{1}{n}\sum_{i \epsilon s_1} z_i, \quad \bar{z}_{s_2} = \frac{1}{m}\sum_{i \epsilon s_2} z_i, \quad \bar{x}_{s_1} = \frac{1}{n}\sum_{i \epsilon s_1} x_i, \quad \bar{x}_{s_2} = \frac{1}{m}\sum_{i \epsilon s_2} x_i$$

$$\text{and} \quad \bar{y}_{s_2} = \frac{1}{m}\sum_{i \epsilon s_2} y_i.$$

Also

$$R_{xz} = \frac{\bar{X}}{\bar{Z}}, \quad R_{yz} = \frac{\bar{Y}}{\bar{Z}}$$

and R_{xy} is as defined in Chapter 4. A regression-thru-ratio estimator of the population mean of the variate y will be

$$\hat{\theta}_2^{(reg)} = \frac{\bar{y}_{s_2}}{\bar{z}_{s_2}}\bar{Z} - b_{xy}^{(s_2)}\left(\frac{\bar{x}_{s_2}}{\bar{z}_{s_2}}\bar{z}_{s_1} - \frac{\bar{x}_{s_1}}{\bar{z}_{s_1}}\bar{Z}\right), \tag{5.1}$$

where $b_{xy}^{(s_2)}$ is the sample regression coefficient between x and y using sample s_2. To derive an approximate expression for the MSE of $\hat{\theta}_2^{(reg)}$, first we define

$$\tilde{\theta}_2^{(reg)} = \frac{\bar{y}_{s_2}}{\bar{z}_{s_2}}\bar{Z} - \beta_{xy}\left(\frac{\bar{x}_{s_2}}{\bar{z}_{s_2}}\bar{z}_{s_1} - \frac{\bar{x}_{s_1}}{\bar{z}_{s_1}}\bar{Z}\right),$$

where β_{xy} is the population regression coefficient between X and Y. For sufficiently large m,

$$M(\hat{\theta}_2^{(reg)}) \approx V(\tilde{\theta}_2^{(reg)}),$$

since $b_{xy}^{(s_2)} - \beta_{xy} = O_p(1/m)$. Hence

$$M(\hat{\theta}_2^{(reg)}) \approx V(\tilde{\theta}_2^{(reg)}) = VE\left(\tilde{\theta}_2^{(reg)}|s_1\right) + EV\left(\tilde{\theta}_2^{(reg)}|s_1\right), \tag{5.2}$$

where the terms are determined as follows. We first note that $\bar{z}_{s_1}$ converges to $\bar{Z}$ in probability which we denote by $(\bar{z}_{s_1} \rightarrow_p \bar{Z})$ for sufficiently large n. Thus the first term of the above equation to a first order of approximation (i.e. $O(1/n)$) becomes

$$\begin{aligned} VE\left(\tilde{\theta}_2^{(reg)}|s_1\right) &\doteq V\left(\frac{\bar{y}_{s_1}}{\bar{z}_{s_1}}\bar{Z}\right) \\ &= \frac{1}{n}\left[S_y^2 - 2R_{yz}S_{yz} + R_{yz}^2 S_z^2\right]. \end{aligned} \tag{5.3}$$

For the second term, we note that upto order $O_(1/m)$,

$$\begin{aligned} V\left(\tilde{\theta}_2^{(reg)}|s_1\right) &\doteq V\left[\frac{\bar{y}_{s_2} - \beta_{xy}\bar{x}_{s_2}}{\bar{z}_{s_2}}\bar{Z}\right] \\ &= V\left(\frac{\bar{u}_{s_2}}{\bar{z}_{s_2}}\bar{Z}\right) \\ &= \left(\frac{1}{m} - \frac{1}{n}\right)\left[S_{nu}^2 - 2r^* S_{nuz} + r^{*2}S_{nz}^2\right], \end{aligned}$$

where

$$\bar{u}_{s_2} = \bar{y}_{s_2} - \beta_{xy}\bar{x}_{s_2}, \quad r^* = \frac{\bar{y}_{s_1}}{\bar{z}_{s_1}} - \beta_{xy}\frac{\bar{x}_{s_1}}{\bar{z}_{s_1}}, \quad S_{nz}^2 = \frac{1}{n-1}\sum_{i\epsilon s_1}(z_i - \bar{z}_{s_1})^2,$$

$$S_{nu}^2 = \frac{1}{n-1}\sum_{i\epsilon s_1}[(y_i - \beta_{xy}x_i) - (\bar{y}_{s_1} - \beta_{xy}\bar{x}_{s_1})]^2 \text{ and}$$

$$S_{nuz} = \frac{1}{n-1}\sum_{i\epsilon s_1}[(y_i - \beta_{xy}x_i) - (\bar{y}_{s_1} - \beta_{xy}\bar{x}_{s_1})][z_i - \bar{z}_{s_1}].$$

Now, for sufficiently large m, $r^* \rightarrow_p R^*(= \bar{Y}/\bar{Z} - \beta_{xy}\bar{X}/\bar{Z})$, To a first order of approximation the following results hold:

$$E(S_{nu}^2) = V(y_i - \beta_{xy}x_i) = S_y^2 - \frac{S_{xy}^2}{S_x^2},$$

$$E(r^* S_{nuz}) = \left(R_{yz} - \frac{S_{xy}}{S_x^2}R_{xz}\right)\left(S_{yz} - \frac{S_{xy}}{S_x^2}S_{xz}\right),$$

$$E(r^{*2} S_{nz}^2) = \left(R_{yz} - \frac{S_{xy}}{S_x^2}R_{xz}\right)^2 S_z^2.$$

Substituting the above results into the second term of (5.2) we get

$$EV\left(\tilde{\theta}_2^{(reg)}|s_1\right) = \left(\frac{1}{m} - \frac{1}{n}\right)\left\{S_y^2 - \frac{S_{xy}^2}{S_x^2} - 2\left(R_{yz} - \frac{S_{xy}}{S_x^2}R_{xz}\right)\left(S_{yz} - \frac{S_{xy}}{S_x^2}S_{xz}\right) + \left(R_{yz} - \frac{S_{xy}}{S_x^2}R_{xz}\right)^2 S_z^2\right\}. \tag{5.4}$$

The MSE to a first order of approximation of $\hat{\theta}_2^{(reg)}$ (i.e. upto order $0(\frac{1}{m})$) is therefore, obtained by simply adding (5.3) and (5.4) and simplifying we get

$$M\left(\hat{\theta}_2^{(reg)}\right) \doteq V\left(\tilde{\theta}_2^{(reg)}\right) = \frac{1}{n}\left[S_y^2 - 2R_{yz}S_{yz} + R_{yz}^2 S_z^2\right] + \left(\frac{1}{m} - \frac{1}{n}\right)\left\{(1-\rho_{xy})S_y^2 - 2\left(R_{yz} - \frac{S_{xy}}{S_x^2}R_{xz}\right)(\rho_{yz} - \rho_{xy}\rho_{xz})S_y S_z + \left(R_{yz} - \frac{S_{xy}}{S_x^2}R_{xz}\right)^2 S_z^2\right\}. \tag{5.5}$$

In the next two sections we would give a detail discussion of an empirical likelihood regression-thru-ratio and an empirical likelihood regression-thru-regression methods of estimation in two-phase sampling.

5.2 Empirical Likelihood Regression-Thru-Ratio Method of Estimation

An empirical likelihood regression-thru-ratio estimate is given by

$$\hat{\theta}_2^{(E)} = \sum_{i\epsilon s_2} y_i^* p_i - b_{xy}^{(s_2)}\left(\sum_{i\epsilon s_2} x_i^* p_i - \frac{\bar{x}_{s_1}}{\bar{z}_{s_1}}\bar{Z}\right), \tag{5.6}$$

where

$$y_i^* = \frac{y_i}{\bar{z}_{s_1}}\bar{Z} \text{ and } x_i^* = \frac{x_i}{\bar{z}_{s_1}}\bar{Z}.$$

Estimates of p_i are obtained by maximizing

$$\prod_{i \epsilon s_2} p_i$$

subject to

$$\sum_{i \epsilon s_2} p_i = 1 \qquad \sum_{i \epsilon s_2} p_i w_1(x_i, z_i) = 0 \qquad (0 \leq p_i \leq 1). \tag{5.7}$$

where

$$w_1(x_i, z_i) = x_i - \frac{\bar{x}_{s_1}}{\bar{z}_{s_1}} z_i.$$

Using the Lagrange multiplier argument, we find that the solution of (5.7) satisfies

$$p_i = \frac{1}{m\{1 + \lambda w_1(x_i, z_i)\}} \qquad (i \epsilon s_2),$$

$$\sum_{i \epsilon s_2} \frac{w_1(x_i, z_i)}{1 + \lambda w_1(x_i, z_i)} = 0. \tag{5.8}$$

To obtain an approximate expression for the MSE of $\hat{\theta}_2^{(E)}$, first we define

$$\tilde{\theta}_2^{(E)} = \sum_{i \epsilon s_2} y_i^* p_i - \beta_{xy} \left(\sum_{i \epsilon s_2} x_i^* p_i - \frac{\bar{x}_{s_1}}{\bar{z}_{s_1}} \bar{Z} \right).$$

For sufficiently large m,

$$M(\hat{\theta}_2^{(E)}) \approx V(\tilde{\theta}_2^{(E)}),$$

since $b_{xy}^{(s_2)} - \beta_{xy} = O_p(1/m)$. Hence

$$M(\hat{\theta}_2^{(E)}) \approx V(\tilde{\theta}_2^{(E)}) = VE\left(\tilde{\theta}_2^{(E)} | s_1\right) + EV\left(\tilde{\theta}_2^{(E)} | s_1\right). \tag{5.9}$$

The first term of the above equation is analytically equivalent to (5.3). Therefore, we would concentrate on finding an expression for the second term. For sufficiently large m the following result holds. Thus

$$\begin{aligned} V\left(\tilde{\theta}_2^{(E)} | s_1\right) &= V\left[\sum_{i \epsilon s_2} (y_i^* - \beta_{xy} x_i^*) p_i | s_1\right]. \\ &= \left(\frac{1}{m} - \frac{1}{n}\right)\left[\hat{S}_g^2 - \frac{\hat{S}_{gw_1}^2}{\hat{S}_{w_1}^2}\right]. \end{aligned} \tag{5.10}$$

The last equality is obtained simply by appealing to the result of Chen and Qin (1993). We note that

$$\hat{S}_g^2 = \frac{1}{n-1}\sum_{i\epsilon s_1}\left[(y_i^* - \beta_{xy}x_i^*) - \left(\bar{y}_{s_1}^* - \beta_{xy}\bar{x}_{s_1}^*\right)\right]^2,$$

$$\hat{S}_{w_1}^2 = \frac{1}{n-1}\sum_{i\epsilon s_1} w_1^2(x_i, z_i) \text{ and}$$

$$\hat{S}_{gw_1} = \frac{1}{n-1}\sum_{i\epsilon s_1}[(y_i^* - \beta_{xy}x_i^*) - (\bar{y}_{s_1}^* - \beta_{xy}\bar{x}_{s_1}^*)]w_1(x_i, z_i).$$

Since $\bar{Z}/\bar{z}_{s_1} \rightarrow_p 1$, the following results hold to a first order of approximation;

$$E(\hat{S}_g^2) = V(y_i - \beta_{xy}x_i) = S_g^2,$$
$$E(\hat{S}_{w_1}^2) = V(x_i - R_{xz}z_i) = S_{w_1}^2 \text{ and}$$
$$E(\hat{S}_{gw_1}) = Cov[(y_i - \beta_{xy}x_i), (x_i - R_{xz}z_i)] = S_{gw_1}.$$

Again since $\hat{S}_g^2 \rightarrow_p S_g^2$, $\hat{S}_{w_1}^2 \rightarrow_p S_{w_1}^2$ and $\hat{S}_{gw_1} \rightarrow_p S_{gw_1}$ to a first order of approximation we obtain the following,

$$EV\left(\tilde{\theta}_2^{(E)}|s_1\right) = \left(\frac{1}{m} - \frac{1}{n}\right)\left[S_g^2 - \frac{S_{gw}^2}{S_w^2}\right], \tag{5.11}$$

where

$$S_g^2 = V(y_i - \beta_{xy}x_i) = S_y^2 - \frac{S_{xy}^2}{S_x^2},$$
$$S_{w_1}^2 = V(x_i - R_{xz}z_i) = S_x^2 + R_{xz}^2 S_z^2 - 2R_{xz}\rho_{xz}S_xS_z \text{ and}$$
$$S_{gw_1} = Cov[(y_i - \beta_{xy}x_i), (x_i - R_{xz}z_i)] = R_{xz}S_zS_y(\rho_{xz}\rho_{xy} - \rho_{yz}),$$

with ρ_{xz}, ρ_{yz} and ρ_{xy} being the correlation coefficient between x and z, y and z and x and y respectively.

The MSE of $\hat{\theta}_2^{(E)}$ to a first order of approximation (upto $O(1/m)$) is therefore, obtained simply by adding (5.3) and (5.11) giving

$$\begin{aligned} M(\hat{\theta}_2^{(E)}) &\approx V\left(\tilde{\theta}_2^{(E)}\right) \\ &= \frac{1}{n}\left[S_y^2 - 2R_{yz}S_{yz} + R_{yz}^2S_z^2\right] + \left(\frac{1}{m} - \frac{1}{n}\right)\left[S_g^2 - \frac{S_{gw}^2}{S_w^2}\right]. \\ &= \frac{1}{n}\left(S_y^2 - 2R_{yz}S_{yz} + R_{yz}^2S_z^2\right) + \left(\frac{1}{m} - \frac{1}{n}\right)\left\{(1-\rho_{xy}^2)S_y^2\right. \\ &- \left.\frac{R_{xz}^2S_z^2S_y^2(\rho_{xy}\rho_{xz} - \rho_{yz})^2}{S_x^2 - 2R_{xz}S_zS_y\rho_{xz} + R_{xz}^2S_z^2}\right\}. \end{aligned} \tag{5.12}$$

To compare the mean square errors of $\hat{\theta}_2^{(E)}$ and $\hat{\theta}_2^{(reg)}$, it is enough to compare their second terms since their first terms are equivalent. We note that

$$M(\hat{\theta}_2^{(E)}) - M(\hat{\theta}_2^{(reg)}) < 0$$

if

$$\frac{R_{xz}^2 S_z^2 S_y^2 (\rho_{xy}\rho_{xz} - \rho_{yz})^2}{S_x^2 - 2R_{xz}S_z S_y \rho_{xz} + R_{xz}^2 S_z^2} > 2S_y S_x(\rho_{yz} - \rho_{xz}\rho_{xy})(R_{yz} - \rho_{xy}R_{xz}) - (R_{yz} - \rho_{xy}R_{xz})^2 S_z^2.$$

Note that since $\rho_{xy} \geq \rho_{yz}$ and $\rho_{xz} \approx 1$ in most applications, it is easy to see that the above inequality is valid in most application. We illustrate this with the following example.

Example

Suppose $x = z + b$ and $y = z + c$, where b and c are constants, z is an auxiliary variate which is closely related to the main auxiliary variate, x but remotely related to the survey variate y. Suppose $S_z = S_x = S_y = 1$.

We observe that $EV(\tilde{\theta}_2^{(E)}|s_1)$ reduces to

$$EV\left(\tilde{\theta}_2^{(E)}|s_1\right) \doteq 0,$$

and $EV(\tilde{\theta}_2^{(reg)})$ becomes

$$EV\left(\tilde{\theta}_2^{(reg)}|s_1\right) \doteq \left(\frac{1}{m} - \frac{1}{n}\right)\left[R_{yz} - R_{xz}\right]^2,$$

where

$$R_{yz} - R_{xz} = \frac{b-c}{\bar{Z}} \neq 0,$$

unless $b = c$.

5.3 Empirical Likelihood Regression-Thru-Regression Estimation

An empirical likelihood regression-thru-regression estimate is given by

$$\hat{\theta}_2^{(ER)} = \sum_{i \epsilon s_2} y_{lri} p_i, \tag{5.13}$$

where

$$y_{lri} = y_i + b_{yz}^{(s_2)}(\bar{Z} - z_i),$$

and $b_{yz}^{(s_2)}$ is the regression coefficient between y and z using sample s_2. Estimates of $p_i(i\epsilon s_2)$ are again obtained by maximizing

$$\prod_{i\epsilon s_2} p_i$$

subject to

$$\sum_{i\epsilon s_2} p_i = 1, \qquad \sum_{i\epsilon s_2} p_i w_2(x_i, z_i) = 0 \qquad (0 \le p_i \le 1),$$

where $w_2(x_i, z_i) = x_i - \bar{X} + b_{xz}^{(s_2)}(\bar{Z} - z_i)$ and $b_{xz}^{(s_2)}$ is the regression coefficient between x and z using sample s_2. Using Langrange multiplier argument once again, we find that the solution to the above equation satisfies

$$p_i = \frac{1}{m\{1 + \lambda_2 w_2(x_i, z_i)\}} \qquad (i\epsilon s_2),$$

$$\sum_{i\epsilon s_2} \frac{w_2(x_i, z_i)}{1 + \lambda_2 w_2(x_i, z_i)} = 0. \tag{5.14}$$

To derive an approximate expression for the MSE of $\hat{\theta}_2^{(ER)}$, first we define

$$\sum_{i\epsilon s_2} y_{lri} p_i,$$

where

$$\tilde{\theta}_2^{(ER)} = y_{lri} = y_i + \beta_{yz}(\bar{Z} - z_i).$$

For sufficiently large n,

$$M(\hat{\theta}_2^{(ER)}) \approx V(\tilde{\theta}_2^{(ER)}),$$

since $b_{yz}^{(s_2)} - \beta_{yz} = O_p(1/m)$. Hence

$$M(\hat{\theta}_2^{(ER)}) \approx V(\tilde{\theta}_2^{(ER)}) = VE(\tilde{\theta}_2^{(ER)}|s_1) + EV(\tilde{\theta}_2^{(ER)}|s_1).$$

Considering the first term, we see that

$$E(\tilde{\theta}_2^{(ER)}|s_1) = \frac{1}{n}\sum_{i\epsilon s_1} y_{lri} = \bar{y}_n + \beta_{yz}(\bar{Z} - \bar{z}_n).$$

For sufficiently large m, we obtain the following expression

$$VE(\tilde{\theta}_2^{(ER)}|s_1) = \frac{1}{n}(1 - \rho_{yz}^2)S_y^2.$$

For the second term we may appeal to the results of Chen and Qin (1993) and for sufficiently large n we have

$$\hat{S}_{y_{1r}} \rightarrow_p S_{y_{1r}}, \quad \hat{S}_{w_2} \rightarrow_p S_{w_2}, \quad \hat{S}_{y_{1r}w_2} \rightarrow_p S_{y_{1r}w_2}.$$

Also for sufficiently large m, we obtain

$$EV(\tilde{\theta}_2^{(ER)}|s_1) = \left(\frac{1}{m} - \frac{1}{n}\right)\left[S_{y_{1r}}^2 - \frac{S_{y_{1r}w_2}^2}{S_{w_2}^2}\right].$$

We note that

$$S_{y_{1r}}^2 = (1-\rho_{yz}^2)S_y^2, \quad S_{w_2}^2 = (1-\rho_{xz}^2)S_x^2 \text{ and } S_{y_{1r}w_2} = (\rho_{yx} - \rho_{yz}\rho_{xz})S_xS_y.$$

The MSE of $\hat{\theta}_2^{(ER)}$ to $O(\frac{1}{m})$ therefore becomes

$$\begin{aligned} M(\hat{\theta}_2^{(ER)}) &\approx V(\tilde{\theta}_2^{(ER)}) \\ &= \frac{1}{n}(1-\rho_{yz}^2)S_y^2 + \left(\frac{1}{m} - \frac{1}{n}\right)\left[(1-\rho_{yz}^2)S_y^2 - \frac{S_x^2S_y^2(\rho_{xy} - \rho_{yz}\rho_{xz})^2}{1-\rho_{xz}^2}\right]. \end{aligned} \tag{5.15}$$

Suppose $\rho_{xz} \approx 1$, then $\rho_{xy} \approx \rho_{yz}$ then

$$\frac{S_x^2S_y^2(\rho_{xy} - \rho_{yz}\rho_{xz})^2}{1-\rho_{xz}^2} = 0,$$

since 0/0=0, as an analytical result. Therefore

$$M(\hat{\theta}_2^{(ER)}) \approx \frac{1}{m}(1-\rho_{yz}^2)S_y^2.$$

Under this same condition,

$$M(t_2) \approx \frac{1}{m}(1-\rho_{xy}^2)S_y^2.$$

Since $\rho_{xy} \approx \rho_{yz}$, it means that $M(\hat{\theta}_2^{(ER)}) \approx M(t_2)$.

Comparisons of Estimators

We observed in Chapter 4 that the best estimator of the population mean of the survey variate y in two-phase sampling using two auxiliary variables was the regression-cum-regression estimator, t_2, proposed by Sahoo et al. (1994). To compare

our estimators, namely, $\hat{\theta}_2^{(E)}$ and $\hat{\theta}_2^{(ER)}$ to all the other estimators discussed in chapter 4, it is enough for us to demonstrate that our estimator $\hat{\theta}_2^{(ER)}$ which happens to better than $\hat{\theta}_2^{(E)}$, is better than the best estimator t_2 mentioned in Chapter 4. We recall that the MSE of the regression-cum-regression estimator is

$$\begin{aligned} M(t_2) &= M(t_{RG}) - f_{s_1}\rho_{yz}^2 S_y^2 \\ &= S_y^2\left[f_{s_2}(1-\rho_{xy}^2) + f_{s_1}\rho_{xy}^2\right] - f_{s_1}\rho_{yz}^2 S_y^2. \end{aligned}$$

Comparing $M(t_2)$ and $M(\hat{\theta}_2^{(ER)})$, we observe that

$$M(\hat{\theta}_2^{(ER)}) - M(t_2) < 0$$

if

$$\rho_{xy}\rho_{yz}\rho_{xz} \leq \frac{1}{2}\left(\rho_{xy}^2\rho_{xz}^2 + \rho_{yz}^2\right),$$

which is always true. We must also mention that, the empirical likelihood regression-thru-ratio estimator, $\hat{\theta}_2^{(E)}$, that we also proposed performs relatively better than all the other estimators mentioned in Chapter 4. The next section discusses some numerical results.

5.4 Numerical Illustrations

For the purpose of illustration we consider three data sets. The first and second data sets may be found in Singh (1986). The variables of the first data set are described below:

y = area under wheat cultivation in 1979-80,

x = area under wheat cultivation in 1978-79,

z = total cultivated area during 1978-79,

$\rho_{xy} = .978$, $\rho_{yz} = .961$ and $\rho_{xz} = .964$.

The variables of the second data set are described below:

y = area under wheat cultivation in 1974,

x = area under wheat cultivation in 1973,

z = area under wheat cultivation in 1971,

$\rho_{xy} = .930$, $\rho_{yz} = .899$ and $\rho_{xz} = .831$.

The third data set can be found in Cochran (1977). The variables are also described as follows:

y = numbers of inhabitants per city in 1950,

x = numbers of inhabitants per city in 1940,

z = numbers of inhabitants per city in 1930,

$\rho_{xy} = .987$, $\rho_{yz} = .970$ and $\rho_{xz} = .996$.

Relative efficiencies of the estimators proposed in the previous sections compared to mean per unit (direct) estimator $\bar{y}_{s_2}$ are presented in Table 5.1. Relative efficiency of an estimator $E1$ to an estimator $E2$ is defined by

$$REF = \frac{\text{MSE of E2}}{\text{MSE of E1}}.$$

Table 5.1 : Relative Efficiency(REF) of different estimators compared to $\bar{y}_{s_2}$

	Population		
	1	2	3
Estimator	n=16,m=6	n=50,m=20	n=34,m=14
t_1	17.61	25.23	6.14
t_2	17.92	25.61	6.31
$\hat{\theta}_2^{(reg)}$	16.60	22.62	6.51
$\hat{\theta}_2^{(E)}$	17.57	31.60	7.07
$\hat{\theta}_2^{(ER)}$	19.04	39.62	7.82

The results presented in Table 5.1 indicate that the estimator $\hat{\theta}_2^{(ER)}$ performs uniformly better than than their counterparts. We can then say that, as far as estimation is concerned in two-phase sampling using two auxiliary variables, the empirical

likelihood regression-thru-regression estimator, $\hat{\theta}_2^{(ER)}$ is superior to any other known estimator.

5.5 Concluding Remarks and Further Research

Empirical likelihood as introduced by Owen (1988, 1990) has been seen to play an important role in providing inference in nonparametric models. In fact, it parallels the role of likelihood in parametric models in many respects.

In our work, we observed how it may be used in survival data analysis in providing inference to quantiles of a survival distribution which is subject to right censoring at a specified time T_o and also in survey sampling, how it may be used to provide efficient estimates of the population mean of the study variable in two-phase sampling in the presence in the presence of two auxiliary variables.

In obtaining the expression for $\hat{W}$, we simply use the Kaplan-Meier estimate at T_o to complete the empirical likelihood function for the lifetime distribution under study. In a further study, we may consider a different set up of the empirical likelihood function given by

$$\prod_{i=1}^{r} p_i \left[1 - \sum_{i=1}^{r} p_i\right]^{n-r}$$

subject to the constraints

$$\sum_{i=1}^{r} p_i = 1, \ \sum_{i=1}^{n} p_i w(x_i) = 0 \ \ (0 \leq p_i \leq),$$

where $w(x_i) = I_{[x \leq m_X]} - .5$.

We will also investigate the use of empirical likelihood method of estimation in sampling over several occasions in a future study.

Bibliography

[1] Aitchson, J. and Silvey, S. D. (1958). Maximum-likelihood estimation of parameters subject to restraints. *Annals of Mathematical Statistics* **29**, 812–828.

[2] Anderson, J. R., Bernstein, L. and Pike, M. C. (1982). Approximate Confidence intervals for probabilities of Survival and Quantiles in Life–Table Analysis. *Biometrics* **38**, 407–416.

[3] Breslow, N. and Crowley, J. (1974). A large sample study of the life tables and product limit estimates under censorship. *Ann. Statist.* **2**, 437–453.

[4] Brookmeyer, R. and Crowley, J. (1982). A confidence interval for the median Survival Time. *Biometrics* **38**, 29–41.

[5] Brown, B. W., Jr., Hollander, M. and Korwar, R. M. (1974). Nonparametric tests of independence for censored data, with applications to heart transplant studies. In *Reliability and Biometrics Statistical Analysis of Lifelength*, F. Proschan and R. J.

[6] De Wet, J. and Randles, R. H. (1987). On the effect of substituting parameter estimators in limiting χ^2 U and V statistics. *The Annals of Statistics* **15**, 812–828.

[7] Chen, J. and Qin, J. (1993). Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. *Biometrika* **80**, 107–116.

[8] Cochran, W. G. (1977). *Sampling Techniques*, 3rd ed. New York: Wiley.

[9] DiCiccio, T. J., Field, C. A. and Fraser, D. A. S. (1990). Approximations of marginal tail probabilities and inference for scalar parameters. *Biometrika* **77**, 77–95.

[10] DiCiccio, T. J., Hall, P. and Romano, J. P. (1989). Comparison of parametric and empirical likelihood functions. *Biometrika* **76**, 465–476.

[11] DiCiccio, T. J., Hall, P. and Romano, J. P. (1991). Empirical likelihood is Bartlett correctable. *Ann. Statist.* **19**, 1053–1061.

[12] Efron, B. (1981). Nonparametric Standard errors and confidence intervals (with discussion). *Canadian journal of statistics* **9**,139–172.

[13] Emerson, J. (1982). Nonparametric confidence intervals for the median in the presence of right censoring. *Biometrics* **38**, 17–27.

[14] Gillespie, J. M. and Fisher, L. (1979). Confidence bands for the Kaplan–Meier survival curve estimate. *The Annals for Statistics* **7**, 920–924.

[15] Gross, A. J. and Clark, V. A. (1975). *Survival Distributions: Reliability Applications in the Biomedical Sciences.* New York: Wiley.

[16] Haberman, S. J. (1984). Adjustment by minimum discriminant information. *Ann. Statist.* **12** 971–988.

[15] Hall, P. (1990). Pseudo-likelihood theory for empirical likelihood. *Ann. Statist.* **18**, 121–140.

[16] Hall, P. and La Scala, B. (1990). Methodology and Algorithms of Empirical likelihood. *International Statistical Review* **58**, 109–127.

[17] Hall, W. J. and Wellner, A. J. (1980). Confidence bands for a survival curve from censored data. *Biometrika* **67**, 133-143.

[18] Johnson, N. L. and Kotz, S. (1970). *Continuous Univariate Distributions*, Vols. 1 and 2. Boston, Massachusetts: Houghton Mifflin.

[19] Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* **53**, 457–481.

[20] Kiregyera, B. (1980). A chain ratio-type estimator in finite population double sampling using two auxiliary variables. *Metrika* **27**, 217–223.

[21] Kiregyera, B. (1983). Regression-type estimators using two auxiliary variables and the model of double sampling. *Metrika* **31**, 215–226.

[22] Kuk, A. Y. C. and Mak, T. K. (1989). Median estimation in the presence of auxiliary information. *J. R. Statist. Soc.* **B 51**, 261–269.

[23] Lagakos, S. W. (1979). General right–censoring and its impact on the analysis of survival data. *Biometrics* **35**, 139–156.

[24] Lagakos, S. W. and Reid, N. (1981). Estimating convolutions from partially censored data. *Biometrika* **68**, 113–117.

[25] Lawless, J. F. (1982). *Statistical Models and Methods for Lifetime Data.* New York: Wiley.

[26] Meier, P. (1975). Estimation of a distribution function from incomplete observations. In *Perspective in Probability and Statistics*, Ed. J. Gani, pp. 67-87. London: Academic Press.

[27] Moeschberger, M. L. and Klein, P. J. (1985). A comparison of several methods of estimating the survival function when there is right censoring. *Biometrics* **41**, 253–259.

[28] Murthy, M. N. (1967). *Sampling Theory and Methods.* Statistical Publishing Society. Calcutta: India.

[29] Olkin, I. (1958). Multivariate ratio estimation for finite populations. *Biometrika* **45**, 154–165.

[30] Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75**, 237–249.

[31] Owen, A. B. (1990). Empirical likelihood confidence regions. *Annals of Statistics* **18**, 90–120.

[32] Owen, A. B. (1991). Empirical likelihood for linear models. it Ann. Statist. **19**, 1725-1747.

[33] Peto, R., Pike, M. C., Armitage, P., Breslow, N. E., Cox, D. R., Howard, S. V., Mantel, N., McPherson, K., Peto, J. and Smith, P. G. (1977). Design and Analysis of randomized clinical trials requiring prolonged observations of each patient. II. Analysis and examples. *British Journal of Cancer* **35**, 1–39.

[34] Peterson, A. V., JR. (1977). Expressing the Kaplan–Meier Estimator as a Function of Empirical Subsurvival Function. *American Statistical Assoc.* **72**, 854–858.

[35] Prasad, N. G. N. and Srivenkataramana, T. (1980). Double sampling with PPS selection. *Vignana Bharathi* **6**, 52–58.

[36] Qin, J. and Lawless, J. (1991). Empirical likelihood and General Estimating Equations. *Statistics Technical Report 91–10*, Univ. Waterloo.

[37] Randles, R. H. (1982). On the asymptotic normality of statistics with estimated parameters. *The Annals of Statistics* **10**, 462–474.

[38] Reid, N. (1981). Estimating the median survival time. *Biometrika* **68**, 601–608.

[39] Sahoo, J., Sahoo, L. N. and Mohanty, S. (1994). An alternative Approach to estimation in two-phase sampling using two auxiliary variables. *Biometrika* **36** 293–298.

[40] Serfling, J. R. (1980). *Approximation theorems of mathematical statistics.* New York: Wiley.

[41] Sheehy, A. (1988). Kullback-Leibler constrained estimation of probability measures. Stanford technical report No. 137.

[42] Simon, R. and Lee, Y. J. (1982). Nonparametric confidence limits for survival probabilities and median survival time. *Cancer Treatment Reports* **66**, 37–42.

[43] Small, G. C. and McLeish, L. D. (1989). Projection as a method for increasing sensitivity and eliminating nuisance parameters *Biometrika* **76**, 693–703.

[44] Slud, V. E., Byar, P. D. and Sylvan, B. G. (1984). A comparison of reflected versus test-based confidence intervals for the median survival time, based on censored data. *Biometrics* **40**, 587–600.

[45] Srivastava, S. Rani, Khare, B. B. and Srivastava, S. R. (1988). On generalised chainestimator for ratio and product of two population means using auxiliary characters. *Assam statist. Rev.*, **2**, 21–29.

[46] Srivastava, S. Rani, Khare, B. B. and Srivastava, S. R. (1990). A generalised chain ratio estimator for mean of finite population. *J. Ind. Soc. Agic. Statist.* **42**, 108–117.

[47] Thomas, D. R. and Grunkemeier, G. L. (1975). Confidence interval estimation of survival probabilities for censored data. *J Am. Statist. Assoc.* **70**, 865–871.

[48] Wilks, S. S. (1938). The large–sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Statist.* **9**, 60–62.

[49] Wong, A. C. M. (1992). Converting observed likelihood to levels of significance for transformation models. *Comm. of Statist–Theory and Methods* **21**, 2809–2823.